

A dual confidence evaluation-based shared control approach for human-machine collaboration

Yaqing Zhou^a, Yun-Bo Zhao^{a, b, *}, Pengfei Li^a, Xia Tian^a, Shuyue Jiang^a, Yu Kang^{a, b}

^a Department of Automation, University of Science and Technology of China, Hefei, China

^b Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

ARTICLE INFO

Communicated by S. He

Keywords:

Human-machine collaboration

Shared control

Reinforcement learning

ABSTRACT

Shared control has become a key strategy for enhancing the safety and adaptability of human-machine collaboration systems, particularly in complex and uncertain environments. However, existing rule-based and confidence-based authority allocation approaches often suffer from limited generalizability or excessive reliance on physiological signals, which hinders their practical deployment. This paper proposes a Dual Confidence-Based Shared Control (DC-SC) approach that enables dynamic and interpretable authority allocation by quantifying the decision confidence of both humans and machines. The human confidence model is constructed through a knowledge-task matching function that measures the cognitive alignment between the operator's expertise and task difficulty, while the machine confidence model assesses decision reliability via an uncertainty-tolerance matching mechanism. These two types of confidence indicators are jointly used to construct a shared control policy, in which the fusion weights are dynamically adjusted using environmental feedback within a policy gradient optimization framework, thereby maximizing human-machine collaborative performance. Theoretical analysis validates the soundness of the confidence models, and experiments conducted in benchmark environments such as LunarLander and UAV path planning demonstrate that DC-SC significantly outperforms both reinforcement learning baselines and traditional shared control approaches in terms of policy performance and system safety.

1. Introduction

In recent years, research on human-machine collaboration (HMC) has gained significant attention, aiming to optimize coordination between humans and machines through dynamic authority allocation. Such collaboration enhances both the efficiency and flexibility of decision-making, particularly in complex and uncertain environments. Although machines possess powerful computational capabilities and execute tasks efficiently, they often struggle to adapt to environmental changes or complex decision scenarios in collaborative settings. By introducing shared control mechanisms, human intuition and experience can be effectively integrated with machine intelligence, enabling the system to flexibly adjust control policies to different contexts and ultimately enhance overall performance and adaptability. Currently, shared control has shown great promise in a range of applications, including autonomous driving [1,2], robotic collaboration [3,4], and UAV navigation [5,6].

In the field of shared control, authority allocation is crucial for ensuring efficient collaboration between systems and human operators.

Current research primarily falls into two categories: rule-based and confidence-based authority allocation. Rule-based approaches grant authority based on predefined logic or conditions, such as differential game theory [7], fuzzy control [8], and adaptive adjustment mechanisms [9]. These approaches offer structured solutions for specific scenarios but tend to be sensitive to environmental changes and require parameter tuning, which increases deployment complexity. Confidence-based approaches simplify the process by dynamically adjusting authority according to the decision maker's confidence level. These approaches rely on physiological signals, such as eye-tracking, electroencephalography (EEG), and electrodermal activity (EDA) [10,11], to quantify psychological states, providing both continuity and real-time responsiveness. However, mapping these signals to confidence levels is inherently complex and often lacks interpretability, requiring extensive data support. Moreover, solely focusing on human decision-makers' confidence is insufficient. The confidence associated with machine decisions must also be considered to fully leverage the complementary strengths of human-machine systems. A comprehensive evaluation of both parties' confidence enables a more holistic assessment of action quality, helping

* Corresponding author at: Department of Automation, University of Science and Technology of China, Hefei, China.

Email address: ybzhao@ustc.edu.cn (Y.-B. Zhao).

to determine the optimal decision path or weighting policy. Therefore, future research should focus on developing intelligent algorithms that integrate multi-source information, thereby enhancing the effectiveness and reliability of shared control systems and fostering more adaptive and flexible models of human-machine collaboration.

Inspired by the above discussion, this paper proposes a Dual Confidence-Based Shared Control (DC-SC) approach, which models the decision confidence of both humans and machines to enable dynamic adjustment of the shared control policy. The human confidence model is constructed based on the marginal payoff of knowledge under varying task difficulties, quantifying the operator's cognitive reliability in the current task. The machine confidence model evaluates the marginal risk of policy output through an uncertainty-tolerance matching mechanism. The proposed confidence evaluation approach offers higher interpretability and generalizability, significantly reducing dependence on physiological data and prior experience. Finally, the two types of confidence signals are used to guide the policy gradient-based optimization of fusion weights, enabling the system to adaptively allocate control authority according to task phases and environmental feedback, thereby enhancing the safety and decision-making performance of human-machine collaboration.

We briefly list the contribution points as follows.

1. **A quantitative evaluation model for human and machine decision confidence is proposed.** The intrinsic role of confidence in the decision-making process is revealed for the first time, and its rationality is validated through theoretical analysis, supporting the principled design of shared control policies.
2. **A confidence-aware shared policy optimization mechanism is developed.** The fusion weights are adaptively adjusted via policy gradient learning, establishing a collaborative optimization method aimed at maximizing overall system performance and ensuring effective human-machine cooperation.
3. **The proposed DC-SC algorithm is implemented and validated.** It is integrated within an actor-critic reinforcement learning framework and empirically evaluated through comparative experiments on both discrete and continuous control tasks.

The remainder of the paper is organized as follows: Section 2 introduces the related work; Section 4 outlines the dual confidence based shared control approach; Section 5 analyzes the properties of the proposed "dual confidence" model; Section 6 describes the experimental setup and analyzes the experimental results; and Section 7 concludes the paper.

2. Related work

Human-machine collaboration. In recent years, the rapid development of machine learning has shifted the capabilities of machines from traditional mechanization and automation to intelligence [12]. Although machines excel in information processing and analysis, surpassing human decision-makers in terms of capacity and speed, their adaptability remains limited, especially in highly complex environments [13]. In contrast, humans can leverage intuition, experience, and creative thinking to make flexible judgments in dynamic and uncertain environments. As a result, human-machine collaboration has shown significant potential in various application fields, including medical diagnosis [14], smart manufacturing [15], and computer vision [16]. Under the concept of human-machine systems, the integration of human and machine capabilities has become a key research focus. Literature [17] points out that human-machine systems have the potential to handle complex tasks, especially in unstructured environments. By combining human cognitive abilities with machine intelligence, many advanced applications have been realized, such as remote grasping behavior systems [18], which can take advantage of human-machine collaboration in unknown or complex environments. Additionally, research in human-machine collaboration is also widely reflected in the

field of demonstration learning [19,20], where it explores how machines can perform specific tasks through human demonstrations and guidance. In these applications, the close collaboration between humans and machines enhances the system's flexibility and adaptability, promoting the in-depth development of human-machine joint decision-making.

Shared control. In the framework of shared control [21,22], humans and machines jointly control the system to optimize performance. The fundamental structure of shared control involves arbitration between human and machine inputs [1], which allocates control authority between the two. There are various types of arbitration methods [23–25], with linear arbitration being widely used due to its simplicity, adjustability, and flexibility [24]. The arbitration factor plays a key role in this design, enabling the dynamic adjustment of human and machine involvement in the control process to optimize decisions across different scenarios [26]. However, achieving effective policy blending within the linear arbitration framework still faces numerous challenges, such as diverse task variations and environmental uncertainties. To address this issue, this paper utilizes policy gradient learning to adaptively update the blending weights, thereby designing a dynamic and reliable shared control policy.

Human-in-the-loop reinforcement learning. Human-in-the-loop reinforcement learning aims to enhance the generalization capability of reinforcement learning during task execution by incorporating human prior knowledge and timely intervention. In recent years, this approach has been widely applied in various domains such as robotic control and policy transfer [18,27,28]. Existing studies on human-RL collaborative learning typically assume that humans possess high-performance policies [29,30], and leverage them as critical guidance for RL policy learning. For instance, Singi et al. [31] proposed a model-free RL-based human-machine collaboration framework that guides the agent to request human assistance in critical situations by estimating decision uncertainty. Similarly, Abel et al. [32] introduced a teacher-student paradigm in which human experts are incorporated into the policy learning process to improve efficiency. However, in practical applications, humans often do not possess high-performance policies. Therefore, it is necessary to consider the decision confidence of both humans and machines in order to fully leverage the complementary advantages of human-machine systems. This paper, for the first time, establishes a human confidence evaluation model by mapping the relationship between human knowledge level and decision quality. For measuring machine confidence, we first assess the uncertainty of machine decisions based on Bayesian reinforcement learning [33], and then construct a machine confidence evaluation model by mapping the relationship between uncertainty and decision quality.

3. Preliminaries

In this section, we introduce the fundamental notations and concepts of RL, as well as the actor-critic framework. The two components discussed in this section together form the foundation of the proposed DC-SC algorithm.

3.1. Notation

We consider a standard RL setting in which an RL agent interacts with a controlled environment. This interaction can be described using a Markov Decision Process (MDP), defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$. The state space \mathcal{S} consists of a set of continuous variables s ; the action space $\mathcal{A} = \mathcal{A}^h \cup \mathcal{A}^m$ includes both human and machine actions, which can be either discrete or continuous. The environment generates a state transition probability $\mathcal{T}(\cdot|s, a) : \mathcal{S} \times \mathcal{A} \rightarrow P(\mathcal{S}')$ which maps a state-action pair (s, a) to a probability distribution over the next state; the reward function $\mathcal{R}(\cdot|s, a) : \mathcal{S} \times \mathcal{A} \rightarrow r$ maps the state-action pair (s, a) to a deterministic reward value r .

At each time step t , the agent observes a state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}$ to send to the environment. The environment returns a scalar reward r_t and the next state s_{t+1} . The agent's behavior is determined by a

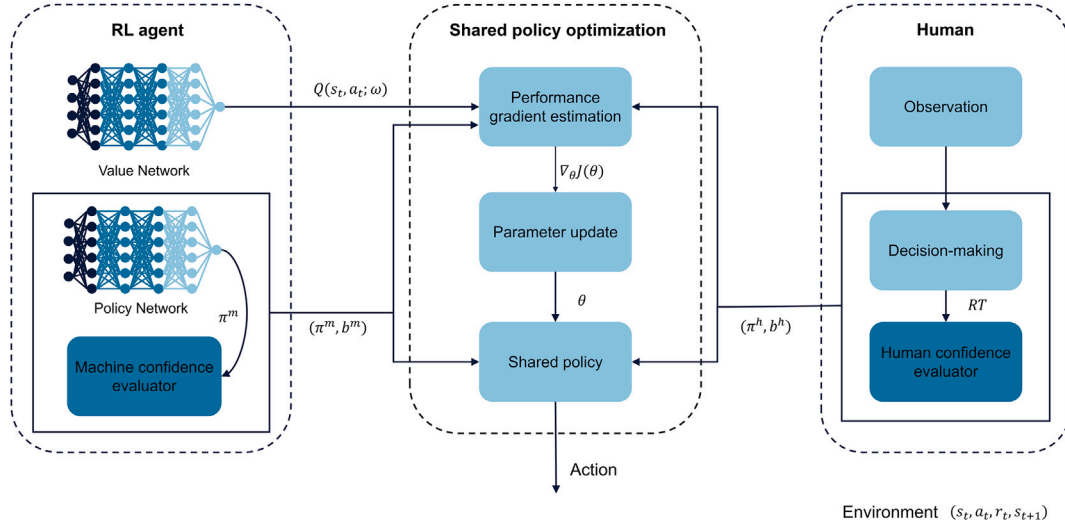


Fig. 1. Dual confidence-based shared control framework for HMC. The framework consists of three core modules: (1) Human module observes the environment and makes decisions, where RT is used to evaluate human confidence b^h ; (2) RL agent module outputs machine policy via a policy network and estimates machine confidence b^m ; (3) Shared policy optimization module fuses human and machine policies and confidence scores to generate the shared policy, and updates fusion parameters via policy gradient methods. The system interacts with the environment in a closed loop using the optimized action.

policy $\pi(a_t|s_t) : S \rightarrow P(a_t)$ which maps the current state to a probability distribution over candidate actions.

3.2. Actor-critic architecture

The objective of RL is to optimize the policy $\pi(a_t|s_t)$ in order to maximize the expected cumulative reward under the dynamics of the environment:

$$J(\theta) = \int_{a_t \sim \pi} \pi Q^\pi(s_t, a_t) da_t \quad (1)$$

where $Q^\pi(s_t, a_t)$ denotes the state-action value function, Under any policy π , Q is defined as follows:

$$Q^\pi(s_t, a_t) = r_t + \gamma \mathbb{E}_{(s_{t+1}, a_{t+1})} [Q^\pi(s_{t+1}, a_{t+1})] \quad (2)$$

where $\gamma \in (0, 1)$ is the discount factor. Then, the policy function can be obtained by maximizing Q :

$$\mu(\cdot|s_t) = \arg \max_{\pi} \mathbb{E}_{(s,a)} [Q^\pi(s,a)] \quad (3)$$

In complex environments, the state transition function \mathcal{T} that characterizes the system dynamics is often unknown. Instead of explicitly modeling \mathcal{T} , neural networks are employed as function approximators to construct the value function and the policy function, with the objectives optimized through loss functions. The value loss \mathcal{L}^Q and the policy loss \mathcal{L}^μ are defined as follows:

$$\mathcal{L}^Q(\omega) = r_t + \gamma \mathbb{E} [Q(s_{t+1}, \mu(s_{t+1}; \phi); \omega)] - Q(s_t, a_t; \omega) \quad (4)$$

and

$$\mathcal{L}^\mu(\phi) = -Q(s_t, \mu(\cdot|s_t; \phi); \omega) \quad (5)$$

Here, $Q(\cdot; \omega)$ denotes the parameterized value function network (also referred to as the critic), and ω represents the parameters of the value network; $\mu(\cdot|s_t; \phi)$ denotes the parameterized policy network (i.e., the actor), and ϕ represents the parameters of the policy network.

4. Dual confidence-based shared control for HMC

In this section, we explore a shared control scheme based on dual confidence modeling, as illustrated in Fig. 1. We first present the evaluation methods for human confidence $b^h(c_t)$ and machine confidence $b^m(c_t)$. These two confidence signals are then jointly incorporated into the fusion weight $\beta_t(\theta)$, based on which a parameterized shared policy $\pi(\cdot|s_t; \beta_t(\theta))$ is constructed. The policy is optimized under the policy gradient framework, enabling dynamic and interpretable authority allocation.

4.1. Human confidence evaluation

This part primarily considers the impact of human knowledge level k_t on human decision confidence $b^h(k_t)$.

Due to the unique learning ability of humans, their knowledge level changes continuously during interactions with the environment. To describe this potential change process, we incorporate response time (RT) as additional information into the model. RT is interpreted as a proxy for cognitive load only under active task engagement and normal vigilance conditions. External factors such as fatigue or passive states (e.g., drowsiness) are not captured by this model and represent boundary conditions for its applicability, and it has been extensively studied in psychology and computational neuroscience [34,35].

RT refers to the time it takes for a human to respond and complete a decision in a task, when RT is within the normal range, humans learn from task feedback, enhancing knowledge acquisition. When RT exceeds the normal range, normal learning cannot take place, leading to a decrease in the amount of knowledge retained, which in turn affects task performance. During the task execution process, the evolution of the knowledge level k_t held by humans is modeled as follows:

$$k_{t+1} = \begin{cases} K & k_t \geq K \\ (1 + e^{-RT})k_t & k_t < K \text{ \& } RT \leq t_d \\ (1 - e^{RT-t_d})k_t & k_t < K \text{ \& } RT > t_d \end{cases} \quad (6)$$

where, t_d represents the upper limit of the normal reaction time (RT). If this time is exceeded, it indicates that the human cognitive load is too high, preventing normal learning. The task difficulty value $K \in (0, \infty) \cup \{\infty\}$ represents the amount of knowledge required for a human to fully learn the task. When the task difficulty value $K = \infty$, it indicates that the

task is impossible to complete because the specifics of the task prevent the human from fully learning it. Additionally, we always assume that the amount of knowledge k_t that the human currently knows is less than or equal to the total knowledge K required to complete the task. For humans, most tasks are easy to complete, but there is always a certain probability that some tasks cannot be completed. This means we assume that the distribution of task difficulty values has a positive probability p at $K = \infty$, and follows a standard exponential distribution for $K \in (0, \infty)$. Specifically, the distribution of task difficulty values can be expressed as follows:

$$\mu(K) = \begin{cases} (1-p)\eta^K \log \frac{1}{\eta} & K > 0. \\ p & K = \infty \end{cases} \quad \eta \in (0, 1) \quad (7)$$

The values k_t and K determine the marginal payoff $f(k_t, K)$. To model decision confidence, we assume that this function has the following characteristics:

1. $f(k_t, \cdot)$ is a decreasing function of task difficulty K , the greater the difficulty, the lower the marginal payoff for humans to reach the same knowledge level k_t ;
2. $f(\cdot, K)$ is an increasing function of the knowledge level k_t , the diminishing returns reflect the positive impact of accumulated knowledge on task completion;
3. When $K = \infty$, $f(k_t, \infty) = 0$ reflecting the characteristics of an impossible task, in this case, the marginal payoff is at its minimum;
4. When $k_t = K$, $f(K, K) = 1$ reflecting that the current knowledge level of humans matches the total knowledge required to complete the task, the marginal payoff reaches its maximum at this point;
5. $f(k_t, K)$ is continuously differentiable, meaning that small changes in the knowledge level or task difficulty will not result in abrupt changes in the marginal payoff;
6. The marginal payoff function is normalized so that its minimum value is 0 and its maximum value is 1: $0 \leq f(k_t, K) \leq 1$.

One example of a marginal payoff function that satisfies the above conditions is:

$$f(k_t, K) = \left(\frac{k_t}{K}\right)^{\delta_1}, \quad \delta_1 > 0 \quad (8)$$

This family of functions is characterized by polynomial growth in k_t . The parameter $\delta_1 > 0$ allows the model to reflect diminishing marginal payoff, meaning that as k_t increases, the contribution of each additional unit of knowledge to performance improvement decreases. In the real world, once the amount of human knowledge exceeds the basic level required to complete a task, the impact of additional knowledge on performance becomes smaller and smaller.

To clarify human decision confidence under the current knowledge level, we construct a human confidence function $b^h(k_t)$, which quantifies the total marginal payoff at the knowledge level k_t as follows:

$$b^h(k_t) = \frac{\int_{K>k_t} f(k_t, K) d\mu(K)}{\int_{K>k_t} d\mu(K)} \quad (9)$$

The integration range $K > k_t$ is considered because we only focus on tasks that require a higher knowledge level. Through normalization by the denominator, we ensure that the confidence function values range between 0 and 1.

4.2. Machine confidence evaluation

This section mainly explores the impact of the uncertainty c_t in machine policy output on the decision confidence $b^m(c_t)$.

Machine agents trained with RL typically provide black-box solutions, outputting decisions for given situations without explanation. If the agent can simultaneously provide an assessment of its decision confidence, the practical value of the system can be greatly enhanced. We leverage research on Bayesian Neural Networks (BNNs) [36] to quantify potential risks in the machine decision-making process and, based on this, evaluate the confidence of the machine's decisions. Unlike conventional neural networks, the weight parameters in BNNs are random variables that follow a probability distribution rather than fixed values. Using the Monte Carlo (MC) Dropout method [37], the uncertainty c_t in the machine policy output can be calculated as follows:

$$\begin{cases} E[\pi^m] \approx \frac{1}{M} \sum_{j=1}^M \mu(\cdot | s_t; \phi_j) \\ E[(\pi^m)^\top \pi^m] \approx \tau^{-1} I + \frac{1}{M} \sum_{j=1}^M (\mu(\cdot | s_t; \phi_j))^\top \mu(\cdot | s_t; \phi_j) \\ c_t = \text{Var}[\pi^m] = E[(\pi^m)^\top \pi^m] - E[\pi^m]^\top E[\pi^m] \end{cases} \quad (10)$$

where, $\mu(\cdot | s_t; \phi_j)$ is the output of the policy network at time t and system state s_t for the j -th sample. τ^{-1} is the observation noise.

In decision-making processes, the greater the uncertainty, the higher the potential risks associated with the decision. To quantify the cost of decision risks, we need to consider the system's risk tolerance, which refers to the level of risk the system is willing to accept when facing uncertainty and potential losses. To achieve this, we introduce an uncertainty threshold $C \in (0, \infty)$ to represent the system's risk tolerance and assume that its value follows a standard exponential distribution:

$$\mu(C) = \lambda^C \log \frac{1}{\lambda} \quad C > 0, \lambda \in (0, 1) \quad (11)$$

The values c_t and C determine the marginal cost $g(c_t, C)$. To model decision confidence, we assume that this function has the following characteristics:

1. As c_t increases, $g(c_t, C)$ rises monotonically, indicating that greater policy uncertainty leads to higher system risk;
2. As C increases, $g(c_t, C)$ decreases monotonically, indicating that greater error tolerance reduces the risk;
3. $g(c_t, C)$ is continuously differentiable, meaning that small changes in uncertainty or tolerance will not cause abrupt changes in cost;
4. The marginal cost function is bounded between 0 and 1: $g(0, C) = 0$ when $c_t = 0$, and $g(C, C) = 1$ when $c_t = C$.

One example of a marginal cost function that satisfies the above conditions is:

$$g(c_t, C) = \frac{c_t^{\delta_2}}{c_t^{\delta_2} + (C - c_t)^{\delta_2}}, \quad \delta_2 > 1 \quad (12)$$

This family of functions is characterized by polynomial growth in c_t . The parameter $\delta_2 > 1$ allows the model to exhibit marginal effects of increasing increments. As the uncertainty c_t increases, each additional unit of uncertainty results in an escalating increase in the potential risk to the system. Once the uncertainty exceeds the specified threshold, the machine's decision-making will no longer be effective.

To clearly determine the confidence in the machine policy under uncertainty, we construct the machine confidence function $b^m(c_t)$, which quantifies the total risk under the current decision uncertainty c_t :

$$b^m(c_t) = 1 - \frac{\int_{C>c_t} g(c_t, C) d\mu(C)}{\int_{C>c_t} d\mu(C)} \quad (13)$$

The integral range $C > c_t$ indicates that the current uncertainty is within the system's tolerance, and the system can accept the decision. Through normalization by the denominator, we ensure that the confidence function values range between 0 and 1.

4.3. Dual confidence-based shared policy optimization

To achieve efficient decision fusion in human-machine collaboration, this section proposes a dual confidence-based shared policy optimization method. The proposed approach constructs a shared policy by integrating the confidence levels of both human and machine agents, enabling dynamic adjustment of authority allocation.

A learnable parameter θ is introduced as a modulation factor for the shared policy and is iteratively optimized within the policy gradient framework, ensuring performance-optimal human-machine collaboration across different task phases and environmental conditions. Referring to the fusion form of the human-machine shared policy in [18,38], we construct the shared policy $\pi(\cdot | s_t; \beta_t(\theta))$ as a weighted combination of the human policy $\pi^h(\cdot | s_t)$ and the machine policy $\pi^m(\cdot | s_t)$, expressed as:

$$\pi(\cdot | s_t; \beta_t(\theta)) = \beta_t(\theta) \cdot \pi^m(\cdot | s_t) + (1 - \beta_t(\theta)) \cdot \pi^h(\cdot | s_t) \quad (14)$$

The fusion weight $\beta_t(\theta)$ is defined using the following function:

$$\beta_t(\theta) = \frac{1}{1 + \exp(-\theta_t \cdot \frac{b^m(c_t)}{b^m(c_t) + b^h(k_t)})} \quad (15)$$

Here, the learnable parameter θ_t controls the sensitivity of the fusion weight to the confidence ratio. Under the shared policy $\pi(\cdot | s_t; \beta_t(\theta))$, the optimization objective (1) is written as:

$$J(\theta) = \int_{a_t \sim \pi_\theta} \pi_\theta Q^{\pi_\theta}(s_t, a_t) da_t \quad (16)$$

where π_θ is a compact representation of $\pi(\cdot | s_t; \beta_t(\theta))$. To optimize the shared policy π_θ , the parameters θ are updated using gradient ascent:

$$\theta \leftarrow \theta + \alpha \cdot \nabla_\theta J(\theta) \quad (17)$$

Denote:

$$b_t = \frac{b^m(c_t)}{b^m(c_t) + b^h(k_t)} \in [0, 1] \quad (18)$$

According to the policy gradient theorem and the chain rule, we have:

$$\nabla_\theta J(\theta) = b_t (\beta_t(\theta) - \beta_t^2(\theta)) (\pi^m - \pi^h) \cdot \int_{a_t \sim \pi_\theta} Q^{\pi_\theta}(s_t, a_t) da_t \quad (19)$$

The confidence levels of human and machine agents influence the computation of the policy gradient through the confidence ratio b_t , thereby modulating the update magnitude of the parameter θ . Specifically, b_t determines both the sensitivity of the fusion weight $\beta_t(\theta)$ to changes in θ and acts as a scaling factor for the gradient $\nabla_\theta J(\theta)$. When the machine exhibits higher confidence than the human, a larger b_t amplifies the gradient update, driving θ to adjust more rapidly toward favoring machine control, thus making the shared policy more reliant on the machine. Conversely, when the human is more confident, the gradient is suppressed, and the policy tends to rely more on human input. In this process, the expected return $J(\theta)$, serves as the objective function that guides the optimization of the fusion parameter θ . By performing policy gradient ascent, maximizing $J(\theta)$ effectively steers θ toward favoring the agent (human or machine) with higher expected long-term utility, thereby enabling the shared policy $\pi(\cdot | s_t; \beta_t(\theta))$ to dynamically adjust and achieve performance-optimal collaboration.

We adopt actor-critic as the foundation for human-machine shared control. Therefore, our framework can be integrated into actor-critic variants such as A2C [39], DDPG [40], and similar approaches. During the task execution phase, the actor network outputs the machine policy $\pi_m = \mu(\cdot | s; \phi)$, with its parameters ϕ remaining fixed. The critic network $Q(s, a; \omega)$ continuously interacts with the environment and adapts to environmental changes by updating its parameters ω . Meanwhile, the parameters θ of the shared policy are iteratively updated within the policy gradient framework, enabling dynamic allocation of control

Algorithm 1: DC-SC Algorithm.

Input: The value network $Q(s, a; \omega)$ and policy network $\mu(\cdot | s; \phi)$, which network parameters ω and ϕ were trained based on the actor-critic algorithm, learning factor α and β , total time steps T .

- 1 **Initialize:** Shared policy parameters $\theta \in \mathbb{R}^m$.
- 2 **for** $t = 1 \rightarrow T$ **do**
- 3 Receptive human policy π^h , human reaction time and machine policy $\pi^m = \mu(\cdot | s_t; \phi)$;
- 4 Evaluate human knowledge level k_t from Eq. (6) and evaluate machine uncertainty c_t from Eq. (10);
- 5 Evaluate human confidence $b^h(k_t)$ from Eq. (9) and evaluate machine confidence $b^m(c_t)$ from Eq. (13);
- 6 Generate shared policy $\pi(\cdot | s_t; \beta_t(\theta))$ by Eq. (14), where fusion weight $\beta_t(\theta)$ is computed via Eq. (15);
- 7 Sample action a_t from shared policy and execute it, then environment provides new state s_{t+1} and reward r_t ;
- 8 **Update value network ω :**
- 9 Compute current value: $q = Q(s_t, a_t; \omega)$;
- 10 Compute gradient: $d = \frac{\partial}{\partial \omega} Q(s_t, a_t; \omega)$;
- 11 Compute temporal difference target:
 $y = r_t + \gamma \max_a Q(s_{t+1}, a; \omega)$;
- 12 Update: $\omega_{t+1} \leftarrow \omega_t - \beta (q - y) d$;
- 13 **Update shared policy parameter θ :**
- 14 Compute gradient $\nabla_\theta J(\theta)$ by Eq. (19);
- 15 Update: $\theta_{t+1} \leftarrow \theta_t + \alpha \nabla_\theta J(\theta)$.
- 16 **end**

authority between human and machine to cope with varying environmental conditions. The overall algorithm is illustrated in Algorithm 1.

5. Properties of dual confidence model

In this section, we further illustrate the rationality of the proposed human and machine confidence models by analyzing their properties derived from Eqs. (9) and (13).

Although the construction of the shared policy explicitly relies only on the numerical outputs of human and machine confidence, the behavioral trends of these outputs—such as monotonicity, convergence, and their associations with knowledge level and decision uncertainty—directly affect the variation of the fusion weight $\beta_t(\theta)$, which in turn impacts the stability and predictability of dynamic control authority allocation. Therefore, the analysis of the confidence model properties not only provides a theoretical validation of the model's reliability but also serves as a theoretical foundation for the rational design of the shared control policy. Moreover, the properties analyzed in this section are consistent with the confidence evolution trends observed in subsequent experiments, further reinforcing the alignment and credibility between the model design and real-world decision-making behavior.

Lemma 1. The function $f(k_t, K)$ satisfies the condition that as $k_t \rightarrow \infty$, $\int_{K > k_t} \frac{\partial}{\partial k_t} f(k_t, K) \eta^K dK \ll \eta^{k_t}$.

Remark 1. $F(k_t) \ll G(k_t)$: As $k_t \rightarrow \infty$, $\frac{F(k_t)}{G(k_t)} \rightarrow 0$.

Proof.

$$\begin{aligned} 0 &\leq \frac{\int_{K > k_t} \frac{\partial}{\partial k_t} f(k_t, K) \eta^K dK}{\eta^{k_t}} \\ &= \frac{\delta_1 k_t^{\delta_1 - 1} \int_{k_t}^{\infty} (K)^{-\delta_1} \eta^K dK}{\eta^{k_t}} \end{aligned}$$

$$\begin{aligned} &\leq \frac{\delta_1 k_t^{\delta_1-1} k_t^{-\delta_1} \int_{k_t}^{\infty} \eta^K dK}{\eta^{k_t}} \\ &= \frac{\delta_1 \left[\eta^K \frac{1}{\log \eta} \Big|_{k_t}^{\infty} \right]}{k_t \eta^{k_t}} = \frac{\delta_1}{k_t \log \frac{1}{\eta}} \xrightarrow{k_t \rightarrow \infty} 0 \end{aligned}$$

□

Property 1. As $k_t \rightarrow \infty$, $b^h(k_t) \rightarrow 0$.

Proof. Note $\rho(K) = \frac{\mu(K)}{\int_{K>k_t} d\mu(K)}$ we have: $b^h(k_t) = \int_{K>k_t} f(k_t, K) d\rho(K)$ firstly:

$$\begin{aligned} \int_{K>k_t} d\mu(K) &= \int_{K>k_t} (1-p)\eta^K \log \frac{1}{\eta} dK + p \\ &= (1-p) \left[-\eta^K \Big|_{k_t}^{\infty} \right] + p \\ &= (1-p)\eta^{k_t} + p \end{aligned}$$

$$\text{next: } \rho(\infty) = \frac{\mu(\infty)}{\int_{K>k_t} d\mu(K)} = \frac{p}{(1-p)\eta^{k_t} + p} \xrightarrow{k_t \rightarrow \infty} 1$$

finally, by the squeeze theorem:

$$\begin{aligned} 0 &\leq \lim_{k_t \rightarrow \infty} b^h(k_t) \\ &= \lim_{k_t \rightarrow \infty} \left(\int_{K \in (k_t, \infty)} f(k_t, K) d\rho(K) + f(k_t, \infty) \rho(\infty) \right) \\ &= \lim_{k_t \rightarrow \infty} \int_{k_t}^{\infty} f(k_t, K) d\rho(K) \\ &\leq \lim_{k_t \rightarrow \infty} \int_{k_t}^{\infty} d\rho(K) = 0 \end{aligned}$$

The above proof uses the facts that $f(k_t, \infty) = 0$, $f(k_t, K) \leq 1$ and the integral value is 0 when the integral interval is 0. □

Property 2. As $k_t \rightarrow \infty$, $\frac{d}{dk_t} b^h(k_t) < 0$.

Proof.

$$\begin{aligned} \frac{d}{dk_t} b^h(k_t) &= \frac{d}{dk_t} \frac{\int_{K>k_t} f(k_t, K) d\mu(K)}{\int_{K>k_t} d\mu(K)} \\ &= \frac{d}{dk_t} \frac{\int_{k_t}^{\infty} f(k_t, K) d\mu(K) + p f(k_t, \infty)}{p + (1-p)\eta^{k_t}} \\ &= \frac{d}{dk_t} \frac{(1-p) \log \frac{1}{\eta} \int_{k_t}^{\infty} f(k_t, K) \eta^K d\mu(K)}{p + (1-p)\eta^{k_t}} \\ &= \frac{1}{(p + (1-p)\eta^{k_t})^2} \left[(p + (1-p)\eta^{k_t})(1-p) \log \frac{1}{\eta} \right. \\ &\quad \cdot \left(-\eta^{k_t} + \int_{k_t}^{\infty} \frac{\partial}{\partial k_t} f(k_t, K) \eta^K d\mu(K) \right) \\ &\quad \left. - \left((1-p) \log \frac{1}{\eta} \right)^2 \eta^{k_t} \int_{k_t}^{\infty} f(k_t, K) \eta^K d\mu(K) \right] \end{aligned}$$

By Lemma 1, the above final equation is less than 0 as a whole. □

According to Eq. (8), the initial value of human confidence is zero. Combined with Properties 1 and 2, we observe that as knowledge increases, human confidence exhibits a nonlinear pattern of first monotonically increasing and then monotonically decreasing. This is a typical cognitive bias: the quantified form of the Dunning-Kruger effect [41]. Specifically, during task execution, humans continuously learn and accumulate knowledge through interaction with the environment, gradually improving their mastery of the knowledge required to complete the

task. However, task performance does not always improve continuously with increasing knowledge levels. Once a certain level of knowledge is reached, humans may become aware of the limitations of their knowledge, which can hinder their ability to effectively complete the task and even lead to a sharp decline in decision-making performance. This is particularly evident in high-risk decision-making environments, such as aviation, military, and emergency management.

Property 3. The machine confidence function $b^m(c_t)$ is a monotonically decreasing function of decision uncertainty c_t .

Proof. That is, to prove $\frac{d}{dc_t} b^m(c_t) < 0$,

$$\begin{aligned} b^m(c_t) &= 1 - \frac{\log \frac{1}{\lambda} \int_{c_t}^{\infty} g(c_t, C) \lambda^C dC}{\log \frac{1}{\lambda} \int_{c_t}^{\infty} \lambda^C dC} \\ &= 1 - \frac{\log \frac{1}{\lambda} \int_{c_t}^{\infty} g(c_t, C) \lambda^C dC}{\lambda^{c_t}} \\ &= 1 - \log \frac{1}{\lambda} \int_{c_t}^{\infty} g(c_t, C) \lambda^{C-c_t} dC \\ &= 1 - \log \frac{1}{\lambda} \int_0^{\infty} g(c_t, c_t + x) \lambda^x dx \end{aligned}$$

note $\rho(x) = \lambda^x \log \frac{1}{\lambda}$, $x > 0$, obviously, it is an exponential distribution function. We use the Leibniz integral rule to differentiate the integral expression with respect to c_t as follows:

$$\begin{aligned} \frac{d}{dc_t} b^m(c_t) &= -\log \frac{1}{\lambda} \int_0^{\infty} \left(\frac{\partial}{\partial c_t} g(c_t, c_t + x) \right) \lambda^x dx \\ &= -\int_0^{\infty} \frac{\partial}{\partial c_t} g(c_t, c_t + x) d\rho(x) \end{aligned}$$

since $g(c_t, c_t + x)$ is a monotonically increasing function of c_t , we have $\frac{\partial}{\partial c_t} g(c_t, c_t + x) > 0$. As the integrand is positive over the interval $(0, \infty)$, the entire integral is positive, leading to the conclusion that the derivative of $b^m(c_t)$ with respect to c_t is negative. □

Property 3 indicates that, under the given assumptions, as the decision uncertainty c_t increases, the confidence in the decision decreases. This aligns with our intuition that higher uncertainty is associated with lower confidence in the decision outcome.

6. Experiments

In this section, to validate the effectiveness of the proposed DC-SC approach, we conducted experiments in both discrete and continuous task scenarios. The experimental design focuses on the following two aspects: (1) performance comparison; and (2) analysis of the rationality of the confidence model. Together, these aspects provide a comprehensive evaluation of the DC-SC approach in terms of performance and mechanism.

6.1. Experimental setup

1) **Discrete Actions: LunarLander.** We use the LunarLander simulation environment from OpenAI Gym as the validation scenario for discrete actions. The LunarLander environment is a classic reinforcement learning problem, where the task requires the spacecraft to land steadily in a designated landing zone marked with a flag by controlling two side thrusters and a main engine, while avoiding crashes. In the human-machine collaboration setting, the spacecraft has acquired the necessary knowledge of posture, speed, and angle for landing through pretraining, enabling it to control its descent smoothly. Meanwhile, the human has accurate knowledge of the landing target and can guide the spacecraft's descent trajectory. Therefore, human-machine collaboration enables precise localization and smooth arrival at the target point.

In this scenario, the reward is set such that when the spacecraft crashes, flies out of bounds, comes to rest on the ground, or the time

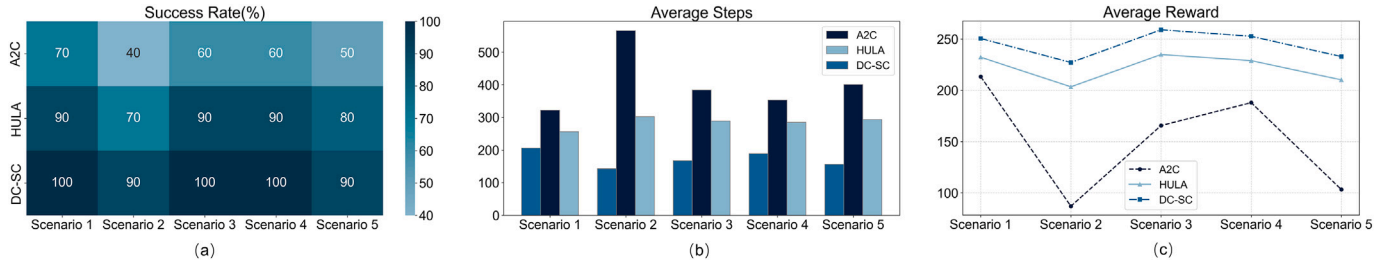


Fig. 2. Performance comparison of the approaches A2C, HULA, and DC-SC in the discrete action scenario (aggregated by every group of ten scenarios). (a) Success rate; (b) Average number of steps; (c) Average cumulative reward.

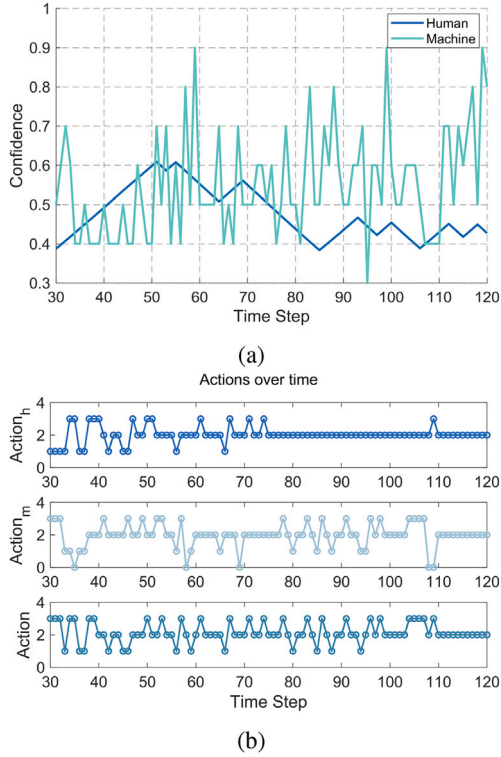


Fig. 3. Authority Allocation of DC-SC in the Discrete Task. (a) Variation in the confidence levels of the machine and the human from step 30 to step 120. (b) Authority allocation from step 30 to step 120: from top to bottom are the machine's actions, the human's actions, and the actions generated by the shared control policy.

runs out, the episode ends, and the agent receives a reward of -100 or 100 depending on whether it lands successfully. Each time one of the spacecraft's legs touches the ground, an additional 10 points are awarded. If the main engine is activated, it consumes fuel and incurs a reward penalty of 0.3 points per frame (assuming unlimited fuel). The action space \mathcal{A} consists of four discrete actions $\{0, 1, 2, 3\}$: 0 means all engines are off, 1 activates the left engine to fire leftward, 2 activates the main engine to fire downward, and 3 activates the right engine to fire rightward. The state $s_t \in S^9$ is a 9-dimensional vector that includes the spacecraft's position $(x(t), y(t))$, velocity $(\dot{x}(t), \dot{y}(t))$, angle and angular velocity $(\theta(t), \dot{\theta}(t))$, contact indicators for the left and right legs with the ground $(leg_l(t), leg_r(t))$, and the coordinates of the landing point $h(t)$.

Before task execution, we train the spacecraft using the A2C [39] algorithm for 500 episodes and use the resulting control policy as the machine policy. The human operator is then allowed to operate the spacecraft alone for 20 runs to become familiar with the controls

and environment, followed by 20 collaborative runs with the trained machine to enable mutual adaptation and optimization. We conduct validation across 50 different scenarios with varying random seeds. The default values of the training hyperparameters used are shown in Table A1. During the experiments, the parameters are set as follows: $p = 0.1, \eta = 0.6, \lambda = 0.6, \delta_1 = 2, \delta_2 = 1.05$.

2) **Continuous Actions: UAV Path Planning.** We build a simulation platform for UAV path planning in uncertain environments on an Ubuntu 20.04 system using a device equipped with a GeForce RTX 3090 GPU and CUDA 11.7. The experimental scenario, proposed in [42], is used as the validation scenario for continuous actions. The objective of the UAV is to avoid obstacles and reach the target point via the shortest possible path. In the human-machine collaboration setting, the UAV has acquired knowledge of shortest path planning through pretraining, while the human can anticipate the trajectories of moving obstacles and guide the UAV to avoid them. Therefore, human-machine collaboration enables safe and optimal arrival at the target point.

In this scenario, the reward is defined such that when the UAV flies out of bounds or collides with an obstacle, it receives a large negative reward. Entering a threat area results in a penalty, which increases as the UAV gets closer to the obstacle. Otherwise, the reward is defined as the inverse function of the distance between the UAV's current position and the target. The action space consists of three types of continuous actions $\{\rho, \sigma, \tau\}$, where $\rho, \sigma \in [0, 3]$ represent the obstacle priority response coefficient and directional response coefficient, respectively, which determine the UAV's movement direction. τ denotes the UAV's heading relative to large obstacles and serves as a heading index, with values in the range $\tau \in [0, 3]$. The state $s_t \in S^9 = \{s_1, s_2, s_3\}$ is a nine-dimensional vector, where s_1 represents the distance between the UAV and the center of the obstacle, s_2 represents the distance between the UAV and the target, and s_3 represents the velocity of the moving obstacle.

Before task execution, we train the UAV using the DDPG [40] algorithm for 500 episodes and use the resulting control policy as the machine policy. The human operator is then allowed to operate the UAV alone for 20 runs to become familiar with the controls and environment, followed by 20 collaborative runs with the trained machine to enable mutual adaptation and optimization. We conduct validation across 50 different scenarios with varying obstacle trajectories and velocities. The default values of the training hyperparameters used are shown in Table A2. During the experiments, the parameters are set as follows: $p = 0.01, \eta = 0.6, \lambda = 0.6, \delta_1 = 2.3, \delta_2 = 1.02$.

In the above two types of tasks, we evaluated the following three approaches: **Typical model-free reinforcement learning approaches (A2C [39] and DDPG [40])**. The machine control policy is trained using RL algorithms and is solely employed for control during the task execution phase; **A popular shared control approach (HULA [31])**. The machine control policy is trained using RL algorithms, and during the task execution phase, human-machine shared control is achieved by evaluating the uncertainty of the machine policy—human intervention is triggered when the uncertainty exceeds a predefined threshold; **A dual**

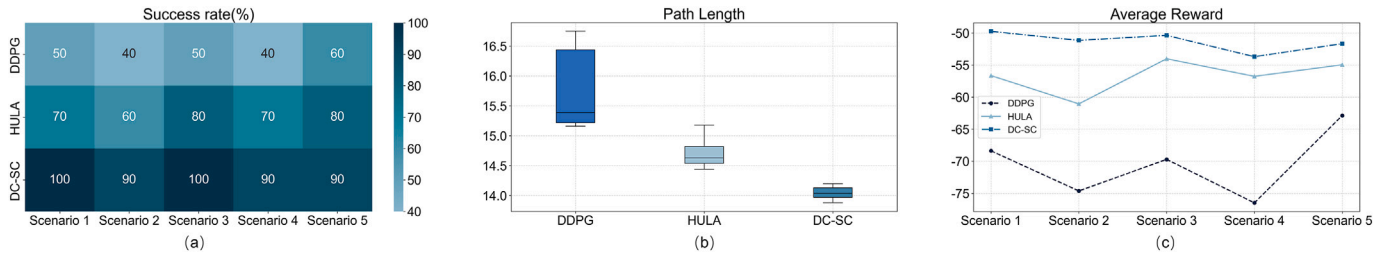


Fig. 4. Performance comparison of the approaches DDPG, HULA, and DC-SC in the continuous action scenario (aggregated by every group of ten scenarios). (a) Success rate; (b) Average path length; (c) Average cumulative reward.

confidence-based shared control approach (DC-SC, the approach proposed in this paper). The machine control policy is trained using RL algorithms, and during the task execution phase, human-machine shared control is achieved by independently evaluating the confidence levels of both machine and human decisions.

To comprehensively evaluate the effectiveness of DC-SC, the following evaluation metrics are established in this experiment: **Performance evaluation metrics.** For the LunarLander task, we use the average cumulative reward, the average number of steps required to complete the task, and the task success rate as performance evaluation metrics. For the UAV path planning task, we use the average cumulative reward, the average path length, and the task success rate as performance evaluation metrics; **Confidence Evolution and Authority Allocation.** To verify the consistency between the actual confidence evolution trends and the theoretical analysis presented in Section 5, as well as the rationality of authority allocation based on confidence evolution, we analyze the temporal changes in human and machine confidence, along with the corresponding variations in human-machine authority allocation over time.

6.2. Experimental results

DC-SC Demonstrates Safer Decision-Making Capability: As shown in Fig. 2(a) and Fig. 4(a), DC-SC consistently maintains a high task success rate across five different scenario groups, significantly outperforming both the standalone machine control approaches A2C/DDPG and the uncertainty-based shared control approach HULA. In most scenarios, DC-SC achieves a success rate of 90% or even 100%, while the success rates of A2C and HULA exhibit noticeable fluctuations. These results indicate that DC-SC possesses safer decision-making capability when facing complex environmental changes and is more effective in enhancing the stability of human-machine collaborative systems.

DC-SC Requires Fewer Decision Steps: As shown in Fig. 2(b) and Fig. 5(a), DC-SC consistently outperforms both A2C/DDPG and HULA in terms of the average number of steps required to complete the task. DC-SC is able to accomplish tasks with fewer decision steps, demonstrating higher execution efficiency and superior path planning capability. These results indicate that DC-SC can achieve task goals more quickly, reduce redundant operations, and improve resource utilization efficiency and decision-making flexibility in complex environments.

DC-SC Achieves Shorter Path Lengths: As shown in the path length boxplot in Fig. 4(b), the paths generated by the DC-SC approach are significantly shorter overall compared to those produced by the other two control approaches across multiple experiments. Additionally, the 3D trajectory visualization in Fig. 5(b) shows that DC-SC can more effectively avoid dynamic obstacles, generating paths that are not only closer to the optimal shortest route but also smoother in trajectory. These results indicate that DC-SC possesses stronger obstacle avoidance and trajectory optimization capabilities in complex environments, enabling more efficient target-reaching while ensuring safety.

DC-SC Achieves Higher Cumulative Rewards: As shown in Fig. 2(c) and Fig. 4(c), DC-SC obtains significantly higher average cumulative

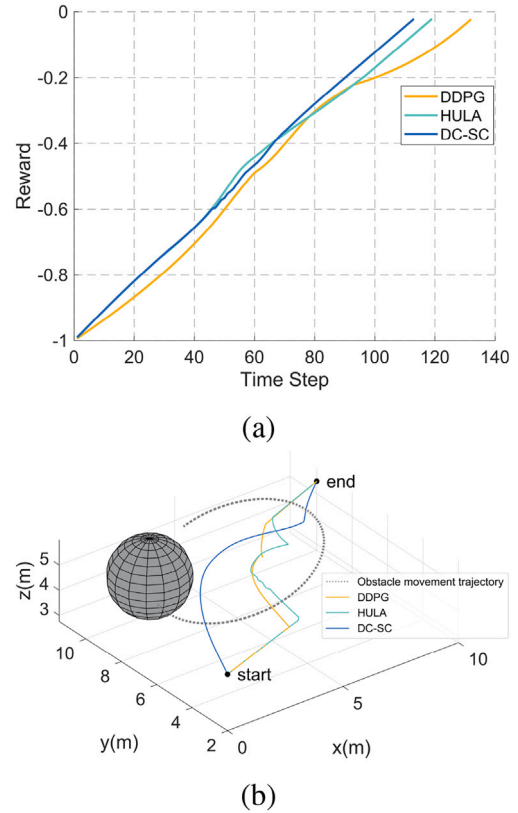


Fig. 5. Performance comparison of DDPG, HULA, and DC-SC. (a) Real-time reward variation. (b) Trajectory variation.

rewards across different task scenarios compared to A2C and HULA, demonstrating superior policy optimization capability. Furthermore, the reward-over-time curves in Fig. 5(a) further validate this finding: DC-SC exhibits a faster reward increase and achieves higher final reward values. These results indicate that DC-SC can more effectively integrate the strengths of both human and machine, enabling superior policy outputs and significantly enhancing overall task performance.

Rationality and Consistency of the Confidence Model: Fig. 3(a) and Fig. 6(a) both show that at the beginning of the task, the human operator has limited understanding of the environment, which may lead to uncertainty and anxiety, resulting in low confidence. As the human interacts with the environment and accumulates experience and knowledge, they gradually develop a more comprehensive understanding of the task, and their confidence increases accordingly. However, in the later stages, increasing workload over time may cause confidence to decline. This non-monotonic trend in human confidence aligns with Properties 1 and 2 in our theoretical analysis and reflects the characteristics of the Dunning-Kruger effect, where confidence evolves with knowledge

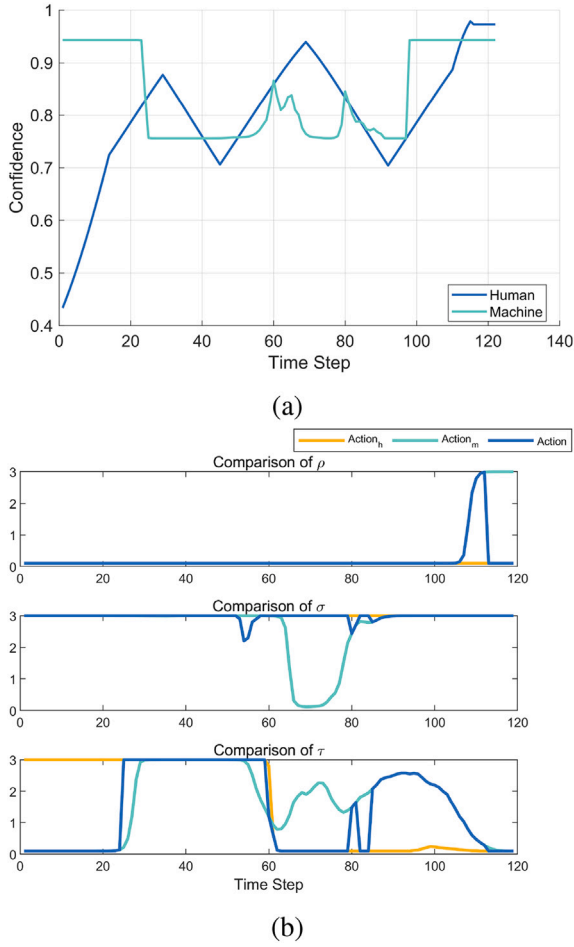


Fig. 6. Authority allocation of DC-SC in the continuous task. (a) Variation in the confidence levels of the machine and the human throughout the entire task. (b) Authority allocation across action dimensions: From top to bottom, it sequentially displays the comparison of action values generated by the machine, the human, and the shared policy in the ρ , σ , and τ dimensions.

accumulation. Meanwhile, because the evolution of knowledge itself is also non-monotonic, human confidence tends to fluctuate, which is more consistent with decision-making intuition in the real world. At the same time, the machine exhibits significant fluctuations during task execution, revealing its limited generalization capability when faced with novel or adversarial scenarios. This observation aligns with [Property 3](#), which states that machine confidence is negatively correlated with decision uncertainty. These results demonstrate that the overall evolution of human and machine confidence not only matches real-world decision-making intuition but also provides empirical support for the theoretical properties of the proposed dual confidence model.

To intuitively illustrate the dynamic process of authority allocation, we plot the variation of control authority over time, as shown in [Fig. 3\(b\)](#) and [Fig. 6\(b\)](#). This variation aligns with the changes in human and machine confidence: when machine confidence is higher, the control authority tends to favor the machine, whereas when human confidence is higher, it tends to favor the human.

7. Conclusion

This paper has focused on the problem of authority allocation in human-machine collaboration and has proposed a DC-SC approach to achieve continuous system performance optimization in dynamic

environments. Specifically, it has first established human and machine decision confidence evaluation models to quantitatively characterize their respective performances in the decision-making process. These confidence measures have then been used to dynamically adjust the relative influence of human and machine decisions. Subsequently, a confidence-driven shared policy has been designed, and within the policy gradient optimization framework, control authority has been adaptively optimized in real time based on environmental feedback to ensure optimal system performance. Furthermore, the theoretical properties of the proposed confidence evaluation models have been analyzed to validate the rationality of the model design. Simulation experiments have been conducted in both discrete and continuous decision-making tasks, and the results have demonstrated the effectiveness of the proposed DC-SC approach. Future research will further focus on shared control approaches under scenarios with inconsistent human and machine goals. It will explore the deep integration of machine intent inference and decision confidence to continuously optimize the overall performance of human-machine collaboration systems while accurately identifying goal intentions.

CRediT authorship contribution statement

Yaqing Zhou: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation. **Yun-Bo Zhao:** Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Pengfei Li:** Project administration, Investigation. **Xia Tian:** Investigation, Data curation. **Shuyue Jiang:** Visualization, Validation, Methodology. **Yu Kang:** Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Table A1

Default values of training hyperparameters used in A2C.

Hyperparameter	Value
Replay buffer size	1×10^5
Batch size	64
Discount factor	0.99
Soft update coefficient	1×10^{-3}
Learning rate	5×10^{-4}
Max grad norm	0.5
Entropy coeff	0.01
Value loss coeff	0.5
Dropout probability	0.01

Table A2

Default values of training hyperparameters used in DDPG.

Hyperparameter	Value
Replay buffer size	1×10^6
Batch size	128
Discount factor	0.99
Critic learning rate	1×10^{-3}
Actor learning rate	1×10^{-4}
Target polyak coefficient	0.995
Initial exploration noise std.	0.3
Grad clip norm	1.0
Dropout probability	0.05

Data availability

Data will be made available on request.

References

- [1] M. Marcano, S. Díaz, J. Pérez, E. Irigoyen, A review of shared control for automated vehicles: theory and applications, *IEEE Trans. Hum.-Mach. Syst.* 50 (6) (2020) 475–491.
- [2] W. Huang, H. Liu, Z. Huang, C. Lv, Safety-aware human-in-the-loop reinforcement learning with shared control for autonomous driving, *IEEE Trans. Intell. Transp. Syst.* 25 (11) (2024) 16181–16192.
- [3] N. Amirshirzad, A. Kumru, E. Oztop, Human adaptation to human–robot shared control, *IEEE Trans. Hum.-Mach. Syst.* 49 (2) (2019) 126–136.
- [4] A. Weiss, A.-K. Wortmeier, B. Kubicek, Cobots in industry 4.0: a roadmap for future practice studies on human–robot collaboration, *IEEE Trans. Hum.-Mach. Syst.* 51 (4) (2021) 335–345.
- [5] S. Zhang, Y. Li, F. Ye, X. Geng, Z. Zhou, T. Shi, A hybrid human-in-the-loop deep reinforcement learning method for UAV motion planning for long trajectories with unpredictable obstacles, *Drones* 7 (5) (2023) 311.
- [6] J. Tan, S. Xue, Z. Guo, H. Li, H. Cao, B. Chen, Data-driven optimal shared control of unmanned aerial vehicles, *Neurocomputing* 622 (2025) 129428.
- [7] M. Flad, L. Fröhlich, S. Hohmann, Cooperative shared control driver assistance systems based on motion primitives and differential games, *IEEE Trans. Hum.-Mach. Syst.* 47 (5) (2017) 711–722.
- [8] Y. Tian, Y. Zhao, Y. Shi, X. Cao, D.-L. Yu, The indirect shared steering control under double loop structure of driver and automation, *IEEE/CAA J. Autom. Sin.* 7 (5) (2020) 1403–1416.
- [9] X. Li, Y. Wang, C. Su, X. Gong, J. Huang, D. Yang, Adaptive authority allocation approach for shared steering control system, *IEEE Trans. Intell. Transp. Syst.* 23 (10) (2022) 19428–19439.
- [10] H. Saeidi, J.D. Opfermann, M. Kam, S. Raghunathan, S. Léonard, A. Krieger, A confidence-based shared control strategy for the smart tissue autonomous robot (star), in: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2018, pp. 1268–1275.
- [11] X. Zhou, P.-C. Liao, Weighing votes in human–machine collaboration for hazard recognition: inferring a hazard-based perceptual threshold and decision confidence from electroencephalogram wavelets, *J. Constr. Eng. Manag.* 149 (9) (2023) 04023084.
- [12] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, *Nature* 529 (7587) (2016) 484–489.
- [13] M.H. Jarrahi, Artificial intelligence and the future of work: human-ai symbiosis in organizational decision making, *Bus. Horiz.* 61 (4) (2018) 577–586.
- [14] W. Hou, Z. Ning, X. Hu, L. Guo, X. Deng, Y. Yang, R.Y.K. Kwok, On-chip hardware accelerator for automated diagnosis through human–machine interactions in healthcare delivery, *IEEE Trans. Autom. Sci. Eng.* 16 (1) (2018) 206–217.
- [15] T. Wang, W. Zhan, Design and control a hybrid human–machine collaborative manufacturing system in operational management technology to enhance human–machine collaboration, *Int. J. Adv. Manuf. Technol.* (2024) 1–11.
- [16] Y. Wang, Y. He, J. Wang, K. Li, L. Sun, J. Yin, M. Zhang, X. Wang, Enhancing intent understanding for ambiguous prompt: a human-machine co-adaptation strategy, *Neurocomputing* (2025) 130415.
- [17] T. Wang, J. Li, Z. Kong, X. Liu, H. Snoussi, H. Lv, Digital Twin improved via visual question answering for vision-language interactive mode in human–machine collaboration, *J. Manuf. Syst.* 58 (2021) 261–269.
- [18] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, Y. Zhu, Robot learning on the job: Human-in-the-loop autonomy and learning during deployment, *arXiv preprint arXiv:2211.08416*, 2022.
- [19] X. Li, X. Wang, X. Zheng, J. Jin, Y. Huang, J.J. Zhang, F.-Y. Wang, Sadrl: merging human experience with machine intelligence via supervised assisted deep reinforcement learning, *Neurocomputing* 467 (2022) 300–309.
- [20] B. Huang, M. Li, R.L. De Souza, J.J. Bryson, A. Billard, A modular approach to learning manipulation strategies from human demonstration, *Auton. Robots* 40 (2016) 903–927.
- [21] D. Gopinath, S. Jain, B.D. Argall, Human-in-the-loop optimization of shared autonomy in assistive Robotics, *IEEE Robot. Autom. Lett.* 2 (1) (2016) 247–254.
- [22] M. Cubuktepe, N. Jansen, M. Alshiekh, U. Topcu, Synthesis of provably correct autonomy protocols for shared control, *IEEE Trans. Autom. Control* 66 (7) (2021) 3251–3258.
- [23] N. Amirshirzad, E. Ugur, O. Bebek, E. Oztop, Adaptive shared control with human intention estimation for human agent collaboration, in: 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE), IEEE, 2022, pp. 653–658.
- [24] G. Maeda, Blending primitive policies in shared control for assisted teleoperation, in: 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 9332–9338.
- [25] Z. Jin, P.R. Pagilla, H. Maske, G. Chowdhary, Task learning, intent prediction, and adaptive blended shared control with application to excavators, *IEEE Trans. Control Syst. Technol.* 29 (1) (2021) 18–28.
- [26] S. Javdani, H. Admoni, S. Pellegrinelli, S.S. Srinivasa, J.A. Bagnell, Shared autonomy via hindsight optimization for teleoperation and teaming, *Int. J. Robot. Res.* 37 (7) (2018) 717–742.
- [27] Q. Li, Z. Peng, H. Wu, L. Feng, B. Zhou, Human-ai shared control via policy dissection, *Adv. Neural Inf. Process. Syst.* 35 (2022) 8853–8867.
- [28] Q. Zhang, Y. Kang, Y.-B. Zhao, P. Li, S. You, Traded control of human–machine systems for sequential decision-making based on reinforcement learning, *IEEE Trans. Artif. Intell.* 3 (4) (2021) 553–566.
- [29] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, S. Srinivasa, Learning from interventions: human-robot interaction as both explicit and implicit feedback, in: 16th Robotics: Science and Systems, RSS 2020, MIT Press Journals, 2020.
- [30] L. Torrey, M. Taylor, Teaching on a budget: agents advising agents in reinforcement learning, in: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-Agent Systems, 2013, pp. 1053–1060.
- [31] S. Singi, Z. He, A. Pan, S. Patel, G.A. Sigurdson, R. Piramuthu, S. Song, M. Ciocarlie, Decision making for human-in-the-loop robotic agents via uncertainty-aware reinforcement learning, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 7939–7945.
- [32] D. Abel, J. Salvatier, A. Stuhlmüller, O. Evans, Agent-agnostic human-in-the-loop reinforcement learning, *arXiv preprint arXiv:1701.04079*, 2017.
- [33] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, Bayesian reinforcement learning: a survey, *Found. Trends Mach. Learn.* 8 (5–6) (2015) 359–483.
- [34] D.R.J. Laming, *Information Theory of Choice-Reaction Times*, Academic Press, 1968.
- [35] S. Sternberg, Memory-scanning: mental processes revealed by reaction-time experiments, *Am. Sci.* 57 (4) (1969) 421–457.
- [36] Y. Gal, Uncertainty in Deep Learning. (Ph.D. thesis), University of Cambridge, 2016.
- [37] Y. Gal, Z. Ghahramani, Dropout as a Bayesian approximation: representing model uncertainty in deep learning, in: International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
- [38] Z. Xue, Z. Peng, Q. Li, Z. Liu, B. Zhou, Guarded policy optimization with imperfect online demonstrations, *arXiv preprint arXiv:2303.01728*, 2023.
- [39] J. Schulman, P. Moritz, S. Levine, M. Jordan, P. Abbeel, High-dimensional continuous control using generalized advantage estimation, *arXiv preprint arXiv:1506.02438*, 2015.
- [40] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, *arXiv preprint arXiv:1509.02971*, 2015.
- [41] D. Dunning, The dunning-kruger effect: on being ignorant of one's own ignorance, in: *Advances in Experimental Social Psychology*, vol. 44, Elsevier, 2011, pp. 247–296.
- [42] P. Yao, H. Wang, Z. Su, Cooperative path planning with applications to target tracking and obstacle avoidance for multi-uavs, *Aerosp. Sci. Technol.* 54 (2016) 10–22.

Author biography



Yaqing Zhou received the BSc degree in mathematics from Hefei University of Technology, Hefei, China, in 2020. She is currently pursuing the Ph.D. degree in the Department of Automation at the University of Science and Technology of China. Her main research interests include AI-driven human-machine autonomous systems and learning based control.



Yun-Bo Zhao received his BSc degree in mathematics from Shandong University, Jinan, China in 2003, MSc degree in systems sciences from the Key Laboratory of Systems and Control, Chinese Academy of Sciences, Beijing, China in 2007, and Ph.D. degree in control engineering from the University of South Wales (formerly University of Glamorgan), Pontypridd, UK in 2008, respectively. He is currently a Professor with University of Science and Technology of China, Hefei, China. He is mainly interested in AI-driven control and automation, specifically, AI-driven networked intelligent control, AI-driven human-machine autonomous systems and AI-driven machine gaming.



Pengfei Li received the Ph.D. degree in control science and engineering from the University of Science and Technology of China, Hefei, China, in 2020. He is an Associate Research Fellow with the School of Information Science and Technology, USTC. His research interests include networked control systems, model predictive control, and learning based control.



Xia Tian received the MSc degree in control science and engineering from the University of Science and Technology of China(USTC), Hefei, China, in 2025. Her main research interests include AI-driven human-machine autonomous systems and learning based control.



Shuyue Jiang received the BSc degree in mathematics from Anhui Normal University, Wuhu, China, in 2020. She is currently pursuing the Ph.D. degree in the Department of Automation at the University of Science and Technology of China. Her main research interests include AI-driven human-machine autonomous systems and learning based control.



Yu Kang received the Ph.D. degree in control theory and control engineering from the University of Science and Technology of China, Hefei, China, in 2005. From 2005 to 2007, he was a Postdoctoral Fellow with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. He is currently a Professor with the Department of Automation and the Institute of Advanced Technology, University of Science and Technology of China. His current research interests include monitoring of vehicle emissions, adaptive/robust control, variable structure control, mobile manipulator, and Markovian jump systems.