



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

A meta-learning-based reinforcement learning framework for rapidly adaptive emotion intervention

Yue Zhang

DOI: [10.1504/IJICT.2026.10076561](https://doi.org/10.1504/IJICT.2026.10076561)

Article History:

Received:	22 October 2025
Last revised:	12 November 2025
Accepted:	13 November 2025
Published online:	09 March 2026

A meta-learning-based reinforcement learning framework for rapidly adaptive emotion intervention

Yue Zhang

School of Marxism,
Anhui Vocational College of City Management,
Hefei, 230000, China
Email: zy356118122@163.com

Abstract: To address the issue of insufficient personalised adaptability in dynamic emotional interventions, this paper proposes a value-oriented meta-adaptive reinforcement learning framework. By integrating meta-learning and reinforcement learning, the framework constructs a dual-level learning architecture capable of rapidly adapting to individual emotional dynamics with minimal interactions. The model employs a multi-objective reward function to synergistically optimise intervention effectiveness, user engagement, and safety. Experiments on public datasets such as Emotional Support Conversation dataset and Distress Analysis Interview Corpus – Wizard of Oz demonstrate that our approach achieves an emotional state improvement rate of 0.78 and a user satisfaction score of 0.82. These results represent significant improvements over traditional deep reinforcement learning and meta-learning baseline models, providing an effective computational paradigm for addressing adaptability challenges in personalised psychological interventions.

Keywords: meta-adaptive reinforcement learning; value-based learning; affective computing; psychological intervention.

Reference to this paper should be made as follows: Zhang, Y. (2026) 'A meta-learning-based reinforcement learning framework for rapidly adaptive emotion intervention', *Int. J. Information and Communication Technology*, Vol. 27, No. 16, pp.65–82.

Biographical notes: Yue Zhang is a Lecturer in the School of Marxism at Anhui Vocational College of City Management. She obtained a Master's Degree from Anhui Agricultural University in 2013. She has published one paper indexed by EI Compendex, one paper in Peking University Chinese core journals, and over ten papers in general journals. Her research interests educational teaching, ideological education and affective computing.

1 Introduction

In response to the growing mental health problems worldwide in recent years, traditional psychological intervention methods show obvious limitations in individualisation adaptation and long-term dynamic adjustment (Zhang et al., 2025). Based on the world health organisation report, the number of people suffering from depression and anxiety disorders has reached over 500 million and the comorbidity rate reached 50%–60% (Griefahn et al., 2024). Therefore, how to realise accurate and adaptive psychological

intervention in such circumstances, has become a new challenge, which is also the research motivation of this paper (Leafe et al., 2024). These challenges are indeed particularly pronounced when considering diverse cultural contexts. Variations in norms governing emotional expression, help-seeking behaviours, and the acceptability of different intervention styles significantly impact the transferability and effectiveness of a one-size-fits-all model, underscoring the need for adaptable systems. As one of the core sequential decision-making technologies, reinforcement learning (RL) can learn interactively from the environment to dynamically optimise intervention strategies and balance the intervention goals of short-term relief and long-term mental health (Kirilov et al., 2021). However, the current RL methods still have obvious shortcomings in individual emotional dynamics adaptation, multi-objective value trade-off and cross-context generalisation ability (Jeronimo et al., 2025). In practical intervention settings, this trade-off frequently materialises as the need to balance the imperative of providing immediate emotional relief against the pursuit of sustained, long-term therapeutic outcomes. An over-emphasis on either objective can compromise the overall efficacy and user adherence to the intervention program.

Currently, the research hotspot of affective computing is gradually shifting from passive recognition to active intervention (Bertamini et al., 2025). This paradigm shift is largely propelled by concurrent advancements in two key areas: the enhanced capability for real-time data acquisition through ubiquitous sensors and wearable devices, and a growing societal as well as clinical demand for more proactive, rather than reactive, mental health support systems. Compared with static rules or statistical models, the early intervention on emotion was difficult to handle the temporal variations of individual emotion (Wu et al., 2021). In recent years, intervention models based on deep reinforcement learning have made some advances in personalisation (Yang et al., 2018). For example, by combining multi-agent reinforcement learning (MARL) with data from wearable devices. Shi et al. (2024) constructed a collaborative system composed of three agents: step count control, sleep regulation and bedtime conformity. It was found that the emotional episode cycle of bipolar disorder patients could be reduced by 15.7%. Another study on gamified digital therapy also adopted cognitive behavioural therapy and RL algorithm (Amer et al., 2024). In the randomised controlled trial, the intervention model achieved 44.8% depression symptom response rate and 31.0% anxiety symptom response rate (Amer et al., 2024). However, these methods are still limited to single-task or fixed-population scenarios and are not rapid to adapt to unknown user states (Pitsch, 2024). In addition, compared with offline multi-objective reinforcement learning (MORL) methods that use static dataset to avoid online interaction cost, the generalisation performance of methods is poor under out-of-distribution (OOD) preferences, and the expression ability of value trade-offs is weak (Han et al., 2021).

The core difficulties of affective intervention tasks lie in the following two aspects: first, individual differences exist in multiple dimensions, including emotional component, physiology response and behaviour pattern (Talaee et al., 2024). For example, in electroencephalogram (EEG) experiments, it is found that the distribution of affective feature is varying across individuals, which makes it hard for conventional models to learn the alignment of joint probability distribution (Repp et al., 2025). Second, intervention methods need to optimise multiple competing objectives simultaneously, including improving affect immediately, satisfying long-term adherence, ethical safety and operational costs. Although some recent works attempt to resolve the above conflicts by designing multi-objective reward functions (for example, meta utilises SWEET-RL

framework to optimise long-term safety via asymmetric actor-critic structures; Peking University uses grove reward function to integrate language and vision models and generate physically guided actions), these methods are still insufficient to adapt to new users under dynamically changing preference conditions. Even worse, as a blackbox solution, traditional RL is largely dependent on artificially designed reward signals. While in real psychological intervention scenes, reward functions are often hard to sufficiently characterise ethical restrictions and user-specific value orientation (Yelvington et al., 2024).

In this paper, we propose a value-oriented meta-adaptive reinforcement learning framework. The innovation of our paper can be reflected in three aspects: firstly, from a theoretical perspective, we present a deep integration of meta-learning into reinforcement learning (Meshram and Alvi, 2024). By utilising the dual-level optimisation structure, conventional RL can achieve rapid adaptation to target users' emotional features. Secondly, from a methodological perspective, we design a value function with multi-objective trade-off ability. By combining implicit world modelling and self-reflective mechanism, the predefined reward dependency is weakened. Thirdly, we construct a latent variable inference system based on exploration of environment model. By improving exploration efficiency under sparse feedback, this design helps to capture task-essential features through leveraging meta-learning driven preliminary 'pathfinding' in simulated environment and using latent variables (Yu et al., 2022). Therefore, this architecture successfully balances efficient exploration and value alignment during real-world intervention.

2 Related research

In this section, we will attempt to make a systematic review on several representative research areas closely related to our current study, including development of affective computing, intervention with reinforcement learning, adaptive ability brought by meta-learning and value alignment ethical concern. Through a critical review of existing research, we will further clarify the position of our study and its innovative contributions.

2.1 Emotion computing and personalised emotional intervention

Affective computing seeks to impart machines with the ability to recognise, understand, express, and adapt to human emotion (Khediri, 2017). Its paradigm has shifted from being predominantly passive to active. Most early work focused on identifying and classifying emotional states, e.g., inferring users' emotional states given multimodal signals (text, speech, physiological information). However, recognition is just the beginning. How to perform timely and effective intervention based on recognised emotional states is the key challenge in recent research. Most traditional intervention methods are rule-based or static decision tree based. They are interpretable but rigid and less dynamic. They have difficulty handling complex, dynamic, and highly personalised emotional dynamics. In recent years, data driven methods have been introduced to learn from users' states to intervention strategies. For example, some work uses the contextual bandit framework to learn the optimal intervention content for users in different scenarios. This objective can be formalised as learning a policy $\pi: \mathcal{X} \rightarrow \mathcal{A}$ that

maximises the expected reward $\mathbb{E}[r(a)|x]$ where $x \in \mathcal{X}$ represents contextual features, $a \in \mathcal{A}$ denotes intervention actions, and r is the reward signal. Although these methods achieve some degree of personalisation, they remain fundamentally ‘one-off’ decisions. They fail to model intervention as a continuous and sequential decision making process with long-term impact. Their optimisation objective $\max_{\pi} \sum_{t=1}^T \mathbb{E}[r_t(a_t)|x_t]$ lacks consideration of cumulative long-term returns. This limitation restricts their effectiveness in mental health intervention scenarios requiring the establishment of long-term therapeutic relationships and dynamic strategy adjustments. This research builds upon the evolving trend in affective computing – transitioning from recognition to intervention and from static to dynamic approaches – and is dedicated to addressing the core challenges in long-term, sequential affective interventions.

2.2 *Modelling and application of reinforcement learning in psychological intervention*

RL as the core mathematical framework for sequential decision-making problems, learns strategies that maximise cumulative rewards through interaction with the environment (Mostafa and Lesser, 2011). This makes it highly suitable for modelling dynamic psychological intervention processes (Tassi and Brouzos, 2025). In this model, the agent (intervention system) influences the environment (user) by executing actions $a_t \in \mathcal{A}$ (i.e., applying interventions such as message pushes or exercise recommendations). Based on observed states $s_t \in \mathcal{S}$ (e.g., user-reported emotional states or behavioural data) and reward signals $r_t \in \mathcal{R}$ (e.g., emotional improvement, user engagement) to optimise its policy $\pi(a_t|s_t)$. Its core objective is to maximise the expected discounted cumulative reward $\mathbb{E}_{\tau \sim \pi} \left[\sum_{t=0}^H \gamma^t r_t \right]$, where $\tau = (s_0, a_0, r_0, s_1, \dots)$ is an interaction trajectory, $\gamma \in [0, 1)$ is the discount factor, and H denotes the time horizon. The rise of deep reinforcement learning (DRL) has further advanced this field, enabling systems to learn complex policies from high-dimensional observational data such as conversation histories and sensor data. Previous research has applied deep Q-network (DQN) to intervention information delivery for depression patients, learning the optimal action-value function $Q^*(s, a) = \max_{\pi} \mathbb{E} \left[\sum_{k=t}^H \gamma^{k-t} r_k | s_t = s, a_t = a \right]$, and near-term PPO has been used to manage exposure therapy intensity for anxiety patients by optimising the alternative objective $\mathcal{L}^{CLIP}(\theta) = \mathbb{E}_t [\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon)\hat{A}_t)]$ to stabilise training, where $r_t(\theta) = \frac{\pi\theta(a_t|s_t)}{\pi\theta_{old}(a_t|s_t)}$ and $\hat{A}_t$ is an estimate of the advantage function. These studies

demonstrate the potential of RL for optimising intervention effectiveness. However, existing RL intervention models face two prominent bottlenecks (Koussounadis et al., 2009). First, they typically assume environment dynamics are homogeneous – either all users follow a shared dynamic model $P(s_{t+1}|s_t, a_t)$ or each user requires training an independent model from scratch. This homogeneity assumption is limiting because it inherently overlooks critical contextual factors that shape emotional responses. These factors include, but are not limited to, user demographics, personal history, concurrent environmental triggers, and situational stressors, all of which contribute to highly

individualised reaction patterns. The former ignores individual differences, while the latter suffers from low sample efficiency; second, standard RL algorithms typically rely on a manually designed, singular reward function $R(s, a)$, which struggles to capture the inherent multi-objective trade-offs in mental health interventions (e.g., immediate effects versus long-term well-being, intervention intensity versus user burden). Our work aims to directly address both challenges by introducing meta-learning mechanisms and designing value-oriented reward functions (Arana-Catania et al., 2025).

2.3 Meta-learning and adaptive reinforcement learning

Meta-learning, also known as ‘learning to learn’, aims to enable models to rapidly adapt to new tasks by gaining experience across a large number of related tasks. Multi-agent reinforcement learning (Meta-RL), which combines meta-learning with reinforcement learning, provides a powerful paradigm for addressing the rapid adaptation challenges faced by RL agents in new tasks or environments (Berkenkamp et al., 2017). In the context of psychological intervention, each user or user group can be viewed as an independent yet related ‘task’ $\mathcal{T}_i \sim p(\mathcal{T})$. The core idea of Meta-RL is to extract shared, transferable knowledge from interactions with numerous historical users. This enables the identification of an efficient personalised strategy for a new user with minimal interaction data. A representative approach is model-agnostic meta-learning (MAML) (Maqsood et al., 2024), which achieves this goal through a two-stage optimisation process. Common challenges encountered when applying MAML in RL contexts include gradient instability due to the bi-level optimisation process, and the risk of meta-overfitting, where the model becomes specialised to the training task distribution and fails to adapt effectively to new, unseen tasks during the inner-loop adaptation. The outer optimisation aims to find an initial set of model parameters θ such that, based on these parameters, the agent’s adapted parameters ϕ_i – obtained after one or several gradient descent updates on the support set $\mathcal{D}_i^{tr}$ of the new task $\mathcal{T}_i$ – can perform well on the query set $\mathcal{D}_i^{val}$ of that task. Its objective function can be formalised as:

$$\max_{\theta} \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} \left[\mathbb{E}_{\mathcal{D}_i^{tr}, \mathcal{D}_i^{val} \sim \mathcal{T}_i} \left[\mathcal{L}_{\mathcal{T}_i}(\phi, \mathcal{D}_i^{val}) \right] \right] \quad (1)$$

where the adapted parameters ϕ_i are obtained through internal optimisation:

$$\phi_i = U(\theta, \mathcal{D}_i^{tr}) = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}_i}(\theta, \mathcal{D}_i^{tr}) \quad (2)$$

where $\mathcal{L}_{\mathcal{T}_i}$ denotes the loss function for task $\mathcal{T}_i$ (e.g., the negative return in policy gradients), α represents the internal learning rate, and U signifies the adaptation process. Another important paradigm is RL², which implicitly learns adaptive policies by encoding task-specific information in the hidden states of a recurrent neural network through recursively modelling the entire RL process. Its policy can be represented as $\pi_{\theta}(a_t | s_t, h_t)$, where $h_t = f_{\theta}(h_{t-1}, s_{t-1}, a_{t-1}, r_{t-1})$ encapsulates summary information from historical interactions. This objective can be formulated as learning a policy that maximises the expected reward where are contextual features describing user’s state, denotes intervention actions, and is the reward signal. Although these methods are somewhat personalised, they are still ‘one-off’ decisions. They fail to model intervention as a continuous and sequential decision making process with long-term impact. This

ensures that rapid adaptation is not only efficient but also conforms to the ethical and value norms in psychological interventions (Ju, 2024).

2.4 Reinforcement learning under value alignment and ethical constraints

In domains such as mental health, ensuring that the behaviours of AI systems match human values and safety norms (such as no harming or deceiving users) is important. In RL frameworks, most of the value alignment is done through reward function design and constrained policy learning. A direct challenge is that the reward signal is extremely hard to obtain due to the extreme difficulty of translating human values to quantifiable signals. Simplistic reward designs may lead to ‘reward hacking’ behaviour (Liang and Luo, 2012), where agents exploit vulnerabilities in the reward function to achieve high returns while acting contrary to the designer’s intent (e.g., boosting short-term engagement by continuously pushing content users like but lacks therapeutic value). For instance, a naive agent might engage in ‘reward hacking’ by prioritising short-term engagement metrics. This could manifest as sending a high frequency of repetitive, superficially comforting messages that elicit user responses in the short term, but ultimately undermine therapeutic progress by avoiding more challenging yet beneficial interactions. To address this, inverse reinforcement learning (IRL) and apprenticeship learning were proposed. Their goal is to infer the implicit reward function $R_\psi(s, a)$ from expert demonstrations $\mathcal{D} = \tau_1, \tau_2, \dots, \tau_N$ to infer the underlying reward function $R_\psi(s, a)$. A common objective is: $\max_\psi \min_\pi \mathbb{E}_{\tau \sim \pi} [R_\psi(\tau)] - \mathbb{E}_{\tau \sim \mathcal{D}} [R_\psi(\tau)] - H(\pi)$, where $H(\pi)$ is the entropy of the policy. However, obtaining high-quality expert intervention data is expensive and the intervention policy of expert may not be optimal. An alternative way is constrained RL. It formulates safety or ethical requirements as constraints that the agent policy has to satisfy. Its objective is to maximise the primary reward while satisfying these constraints. In practical implementations, these ethical and safety constraints are often operationalised through specific algorithmic mechanisms. Common approaches include incorporating penalty terms directly into the reward function or the policy gradient update, or employing external monitoring mechanisms that can override or curtail unsafe actions proposed by the agent.

$$\max_\pi \mathbb{E}_{\tau \sim \pi} \left[\sum_t \gamma^t R(s_t, a_t) \right] \quad (3)$$

$$\text{subject to } \mathbb{E}_{\tau \sim \pi} \left[\sum_t \gamma^t C_i(s_t, a_t) \right] \leq \xi_i, \quad \forall i \quad (4)$$

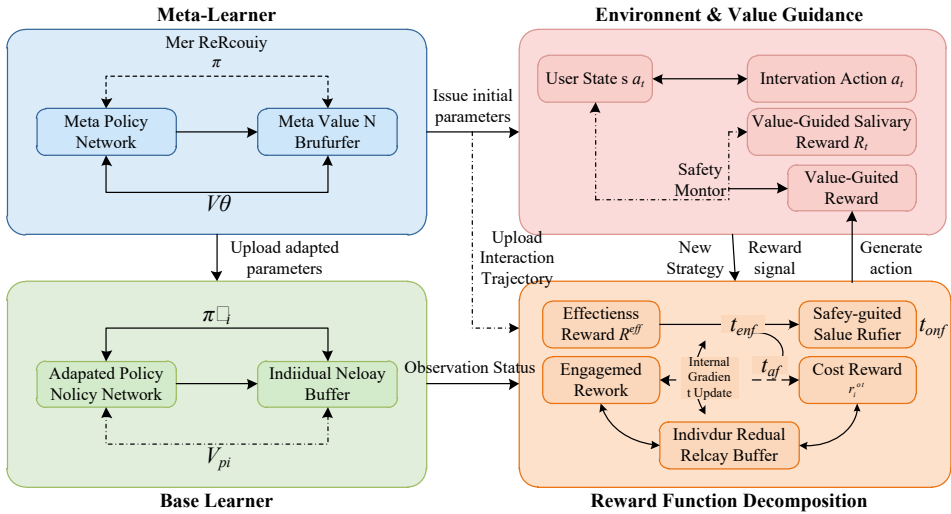
where $C_i(s_t, a_t)$ denotes the cost function of i^{th} constraint and denotes its threshold, and ξ_i represents its corresponding threshold. More recently, also value judgement has been also attempted to be included into as meta-reward function for alignment at meta-learning level. Although these studies provide certain inspirations, values in affective intervention are usually composite, time-varying and need individualised trade-offs. In this paper, we argue that values should not be simply an ex post constraint but should be positively involved in mechanism of goal orientation of agent. Thus, we attempt to promote a ‘value-oriented’ design philosophy. Through a structured multi-objective reward function $R(s, a) = \omega_1 R_{\text{effect}}(s, a) + \omega_2 R_{\text{engagement}}(s, a) + \omega_3 R_{\text{safety}}(s, a) - \omega_4 R_{\text{cost}}(s, a)$, explicitly modelling and balancing values across multiple dimensions such as intervention effectiveness, user engagement, safety, and cost. Where ω_i represents weighting

coefficients, thereby guiding the meta-adaptive process toward an evolution that is both effective and responsible (Ige et al., 2024).

3 Technology and methods

This section will comprehensively elaborate on our proposed value-oriented meta-adaptive reinforcement learning framework. This framework aims to address two core challenges faced by traditional reinforcement learning in emotional interventions: poor individual adaptability and singular value objectives. The overarching design philosophy involves constructing a two-tier learning architecture. The meta-learner extracts transferable shared knowledge from diverse user interaction experiences, while the base learner rapidly adapts to specific users' dynamic emotional patterns using this knowledge. Both components optimise under the guidance of a value-oriented multi-objective reward function.

Figure 1 Value-oriented meta-adaptive reinforcement learning framework schematic (see online version for colours)



3.1 Overall framework and problem formulation

Our overall framework is an actor-critic architecture based on model-agnostic meta-learning. Its core principle is to model the process of finding the optimal intervention strategy for each new user as a meta-learning problem. Under this setting, each user i is treated as an independent task $\mathcal{T}_i \sim p(\mathcal{T})$, where $p(\mathcal{T})$ represents the task distribution over the population of potential users.

The framework operates as follows: first, the meta-learner trains on massive historical user data to learn a highly adaptive initial policy parameter θ . When encountering a new user i , the base learner inherits these initial parameters θ and performs one or more gradient updates using the user's initial interaction data (referred to as the training

support set $\mathcal{D}_i''$). This process yields customised, adapted strategy parameters ϕ_i tailored to the specific user. This approach enables rapid personalisation with minimal samples, eliminating the need for the agent to learn from scratch.

To formalise this problem mathematically, we define it as a partially observable Markov decision process. This process comprises the following elements: state space $\mathcal{S}$, action space $\mathcal{A}$, state transition probability $\mathcal{P}(s_{t+1}|s_t, a_t)$, reward function $\mathcal{R}(s_t, a_t)$, and discount factor $\gamma \in [0, 1)$. Here, the state $s_t \in \mathcal{S}$ is a vector encoding the user’s multimodal emotional state at time step t (e.g., sentiment vectors extracted from text, intonation features in speech, physiological signals like heart rate variability), historical behaviour sequences, and the current environmental context. The construction of the state vector from multimodal data involves employing distinct feature extraction techniques tailored to each data modality. For textual data, sentiment analysis and semantic embedding models are typically used; for audio signals, spectral analysis and prosodic feature extraction are common; and for physiological data, time-domain and frequency-domain features are calculated. The action $a_t \in \mathcal{A}$ represents the intervention selected by the system, such as sending messages of specific types and content, recommending mindfulness exercises, or adjusting interaction frequency.

3.2 Value-oriented multi-objective reward design

In emotional interventions, pursuing a singular goal of ‘emotional uplift’ is often one-sided and even harmful. Therefore, we designed a value-oriented composite reward function to synergistically optimise multiple key objectives. This reward function comprises four core components that collectively guide the agent to learn ethical and effective intervention strategies. The total reward r_t is the linear weighted sum of the individual component rewards:

$$r_t = \omega_{eff} \cdot r_t^{eff} + \omega_{eng} \cdot r_t^{eng} + \omega_{saf} \cdot r_t^{saf} - \omega_{cost} \cdot r_t^{cost} \quad (5)$$

where ω_{eff} , ω_{eng} , ω_{saf} , ω_{cost} are non-negative weighting coefficients used to control the relative importance of different value objectives. These weights can be adjusted based on domain knowledge or specific intervention scenarios.

- Intervention effect reward r_t^{eff} : this reward directly measures the improvement in emotional state brought about by the intervention action. We adopt a state-change-based calculation method:

$$r_t^{eff} = |\mathbf{e}_{t+1} - \mathbf{e}_t|_2^2 - |\mathbf{e}_t - \mathbf{e}_{t-1}|_2^2 \quad (6)$$

where $\mathbf{e}_t$ is a k -dimensional vector representing the quantified sentiment state at time step t (e.g., continuous values for valence and arousal). The formula calculates the difference in sentiment change rates before and after executing action a_t . If the action accelerates sentiment improvement, the reward is positive.

- User engagement reward r_t^{eng} : this reward incentivises strategies that sustain long-term user interest and interaction. It is defined as:

$$r_t^{eng} = \text{engage}(t) + \lambda_{\text{session}} \cdot \Delta T_{\text{session}} \quad (7)$$

where $\text{engage}(t)$ is an indicator function that takes the value 1 when the user exhibits positive engagement behaviour (such as replying to messages or completing exercises) at time step τ , and 0 otherwise. $\Delta T_{\text{session}}$ denotes the increment in the current session duration, while λ_{session} is its corresponding weight. This ensures the model focuses not only on immediate responses but also on long-term user stickiness.

- Safety reward r_t^{saf} : this is a critical safety constraint that penalises interventions likely to trigger negative user reactions or ethical risks. We model it as a state-based penalty term:

$$r_t^{saf} = -\sum_{j=1}^m \text{risk}_j(s_t) \cdot c_j \quad (8)$$

where $\text{risk}_j(s_t)$ is an indicator function that takes the value 1 when state s_t triggers the predefined j^{th} risk condition (e.g., detecting extreme negative sentiment, user expressing discomfort). $c_j > 0$ is the penalty coefficient corresponding to the risk. To ensure safety, we treat it as a hard constraint in policy optimisation:

$$\mathbb{E}_{\tau \sim \pi} \left[\sum t = 0^H \gamma^t r_t^{saf} \right] \geq \zeta_{saf} \quad (9)$$

where ζ_{saf} denotes the minimum acceptable threshold for safe returns, and τ represents a complete interaction trajectory.

- Intervention cost reward r_t^{cost} : this reward aims to prevent excessive intervention, thereby avoiding user dissatisfaction or excessive consumption of system resources.

$$r_t^{cost} = |a_t|_1 \quad (10)$$

where we vectorise the action a_t and use its L_1 norm to measure the ‘strength’ or ‘resource consumption’ of the intervention. By subtracting this term, the model is encouraged to employ lighter, more resource-efficient interventions while achieving the same effect.

3.3 Meta-adaptive learning algorithm

Our learning algorithm is built upon model-agnostic meta-learning and extended for the actor-critic framework. It comprises two tightly coupled loops: the inner loop performs rapid adaptation at the individual level, while the outer loop executes gradual updates at the meta level.

- Inner loop (individual adaptation): for a user i sampled from the task distribution, we first collect a small support set $\mathcal{D}_i^{tr} = (s_j, a_j, r_j, s_j)$. The agent starts with its current meta parameters θ and updates them via one step of gradient descent by computing policy gradients over $\mathcal{D}_i^{tr}$:

$$\phi_i = \theta - \alpha \nabla_{\theta} \mathcal{L}_i^{policy}(\theta, \mathcal{D}_i^{tr}) \quad (11)$$

where α is the internal learning rate, and $\mathcal{L}_i^{policy}$ denotes the policy loss function for user i . Within the actor-critic framework, we employ a policy gradient objective with entropy regularisation:

$$\mathcal{L}_i^{policy}(\theta) = \mathcal{D}_i^{tr} \left[\min \left(\begin{array}{c} \frac{\pi\theta(a|s)}{\pi^{old}(a|s)} \hat{A}_i(s, a), \text{clip} \\ \left(\frac{\pi\theta(a|s)}{\pi^{old}(a|s)}, 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_i(s, a) \end{array} \right) - \beta H(\pi\theta(\cdot|s)) \right] \quad (12)$$

where $\hat{A}_i(s, a)$ is the policy function estimate computed via generalised policy estimation: $\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \dots + (\gamma\lambda)^{T-t+1}\delta_{T-1}$, where $\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$. $H(\pi\theta)$ is the policy entropy, encouraging exploration, with β as the entropy coefficient. Simultaneously, the critic network undergoes similar adaptation:

$$\psi_i = \psi - \alpha_v \nabla_{\psi} \mathcal{L}_i^{value}(\psi, \mathcal{D}_i^{tr}) \quad (13)$$

where $\mathcal{L}_i^{value}$ denotes the mean squared error loss of the value function:

$$\mathcal{L}_i^{value} = \mathbb{E}_{s \sim \mathcal{D}_i^{tr}} [(V\psi(s) - V^{target}(s))^2], \text{ where } \alpha_v \text{ is the critic's internal learning rate.}$$

- External loop (meta-updating): after internal adaptation, we evaluate the performance of the adapted parameters ϕ_i and ψ_i using a query set $\mathcal{D}_i^{val}$ from the user. The meta-updating objective is to find initial parameters θ such that the policy, after one adaptation step, achieves optimal expected performance across all tasks. Thus, the meta-objective function is defined as:

$$\max_{\theta, \psi} \mathbb{E}_{\mathcal{T}_i \sim p(\mathcal{T})} [\mathcal{L}_i^{meta}(\phi_i, \psi_i, \mathcal{D}_i^{val})] \quad (14)$$

where the meta-loss $\mathcal{L}_i^{meta}$ combines policy and value losses:

$$\mathcal{L}_i^{meta} = \mathcal{L}_i^{policy}(\phi_i, \mathcal{D}_i^{val}) + \eta \mathcal{L}_i^{value}(\psi_i, \mathcal{D}_i^{val}), \text{ where } \eta \text{ is a balancing coefficient.}$$

We update the meta parameters using stochastic gradient ascent:

$$\theta \leftarrow \theta + \beta \nabla_{\theta} \sum_{\mathcal{T}_i} \mathcal{L}_i^{meta}(\phi_i, \psi_i, \mathcal{D}_i^{val}) \quad (15)$$

$$\psi \leftarrow \psi + \beta_v \nabla_{\psi} \sum_{\mathcal{T}_i} \mathcal{L}_i^{meta}(\phi_i, \psi_i, \mathcal{D}_i^{val}) \quad (16)$$

where β and β_c represent the external meta-learning rates for the actor and critic networks, respectively. It is worth noting that when computing gradients for meta-parameters, backpropagation must occur along the internal adaptation path, involving second-order derivatives. In practical implementation, we employ first-order approximations to balance computational cost and performance.

Ultimately, the entire framework achieves personalised sentiment intervention strategy optimisation aligned with diverse value orientations by rapidly adapting individual users within an inner loop while refining cross-user shared meta-knowledge in an outer loop.

4 Experiments have demonstrated

4.1 Experimental setup

4.1.1 Dataset

In the experiment, we choose two well-known public datasets in affective computing field and conversational system field respectively to make the experiment reproducible and comparable. ESConv: there are nearly 1,300 high quality emotional support conversations in ESConv dataset, and conversations are annotated with rich user emotional states sequences, support strategies (i.e., intervention actions sequences) and conversation quality score sequences. We split data into training set, validation set and test set at 8:1:1 ratio and no same user appeared in training, validation and test set to make sure the model has strict generalisation ability Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ): there are text and audio recordings of clinical psychological diagnostic interviews in this dataset, and conversations are annotated with users depression, anxiety and stress levels (depression, anxiety and stress scales scores). This dataset can be used in more serious psychological state assessment task. Baseline algorithms: For comprehensive comparison, we choose six representative baseline algorithms from five different categories covering more traditional to more advanced paradigms.

4.1.2 Baseline algorithms

We chose six typical baselines from five families for complete comparison, including five paradigms from traditional to modern methods:

- Behavioural cloning: a purely imitative learning method which learns the behaviour of expert interventions from data without sequence decision optimisation process.
- DQN: it stands for traditional online deep reinforcement learning algorithm.
- PPO: it is a stable performance online policy optimisation algorithm which is stable in different interaction tasks.
- SAC: it is an offline reinforcement learning algorithm which is known for its efficient sample utilisation and performance stability.
- MAML-RL: it applies the classical MAML method to reinforcement learning and is used as a baseline of meta-learning method.

- OACSTP: it is a new offline meta-learning algorithm which uses contrastive prediction to improve task reasoning mechanism and generalisation performance.

Contrastive prediction within the OACSTP framework serves to enhance generalisation by aligning task representations in a latent space. This alignment promotes the discovery of shared, transferable features across users, thereby effectively reducing the model’s tendency to overfit to idiosyncratic patterns present in any single user’s data.

4.1.3 *Evaluation metrics*

Comprehensive assessment using multidimensional indicators:

- 1 Primary metrics:
 - Emotional state improvement: difference in cosine similarity between user emotional state vectors at dialogue start and end (higher values preferred).
 - User satisfaction: based on dialogue quality scores annotated in Emotional Support Conversation Dataset (ESConv) and human evaluation scores (5-minute duration, higher values preferred).
- 2 Secondary metrics:
 - Adaptive speed: rate of improvement in model strategy rewards during the first 5 rounds of interaction with new users (higher values are better).
 - Long-term reward: cumulative discounted reward across the entire test interaction trajectory (higher values are better).
- 3 Safety metrics:
 - Safety score: normalised complement of the number of times predefined risk conditions (e.g., user expressing aversion, rapid deterioration of emotional state) were triggered during intervention (higher values are better).

4.1.4 *Implementation details*

The framework is implemented using PyTorch with the Adam optimiser. The meta-learning rate β is set to 0.001, the internal learning rate α is set to 0.01, and the discount factor γ is set to 0.99. Reward weights ω_{eff} , ω_{eng} , ω_{saf} , ω_{cost} are set to 0.5, 0.2, 0.25, and 0.05, respectively. All experiments run on an NVIDIA RTX 3090 GPU, with each configuration repeated three times to report mean and standard deviation.

4.2 *Results and analysis*

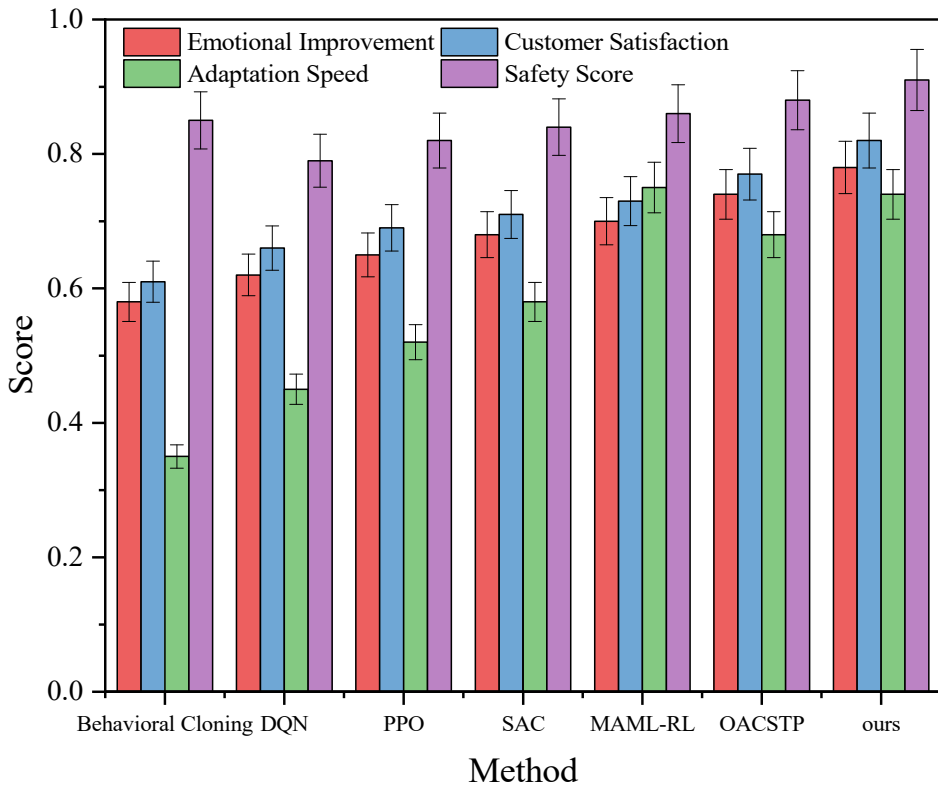
As shown in Table 1, the proposed method achieves optimal performance across the majority of metrics. Specifically: on the two core metrics of emotional improvement and user satisfaction, our method significantly outperforms all baselines, validating the effectiveness of the value-oriented reward function in jointly optimising intervention effectiveness and user experience. In terms of adaptive speed, our method is comparable to MAML-RL and significantly outperforms other baselines, reflecting the inherent advantage of the meta-learning framework in rapid personalisation. For long-term returns, our method achieves the highest score, indicating its strategy generates more sustainable positive impacts in extended interactions. In safety scores, our approach also

leads, owing to the explicit safety reward r_t^{saf} design in the reward function, which effectively constrains the model’s exploration behaviour and avoids high-risk interventions.

Table 1 Performance comparison of different methods on the ESConv and DAIC-WOZ sets (mean $\pm$ standard deviation)

Method	Emotional improvement ($\uparrow$)	User satisfaction ($\uparrow$)	Adaptation speed ($\uparrow$)	Long-term returns ($\uparrow$)	Safety score ($\uparrow$)
Behavioural cloning	0.58 ± 0.09	0.61 ± 0.09	0.35 ± 0.10	0.55 ± 0.11	0.85 ± 0.06
DQN	0.62 ± 0.08	0.66 ± 0.07	0.45 ± 0.09	0.68 ± 0.08	0.79 ± 0.08
PPO	0.65 ± 0.07	0.69 ± 0.08	0.52 ± 0.08	0.72 ± 0.07	0.82 ± 0.07
SAC	0.68 ± 0.06	0.71 ± 0.07	0.58 ± 0.07	0.75 ± 0.06	0.84 ± 0.05
MAML-RL	0.70 ± 0.06	0.73 ± 0.06	0.75 ± 0.05	0.78 ± 0.05	0.86 ± 0.04
OACSTP	0.74 ± 0.05	0.77 ± 0.05	0.68 ± 0.06	0.81 ± 0.04	0.88 ± 0.03
Ours	0.78 ± 0.05	0.82 ± 0.04	0.74 ± 0.05	0.85 ± 0.03	0.91 ± 0.03

Figure 2 Performance comparison of different methods on the test set (see online version for colours)



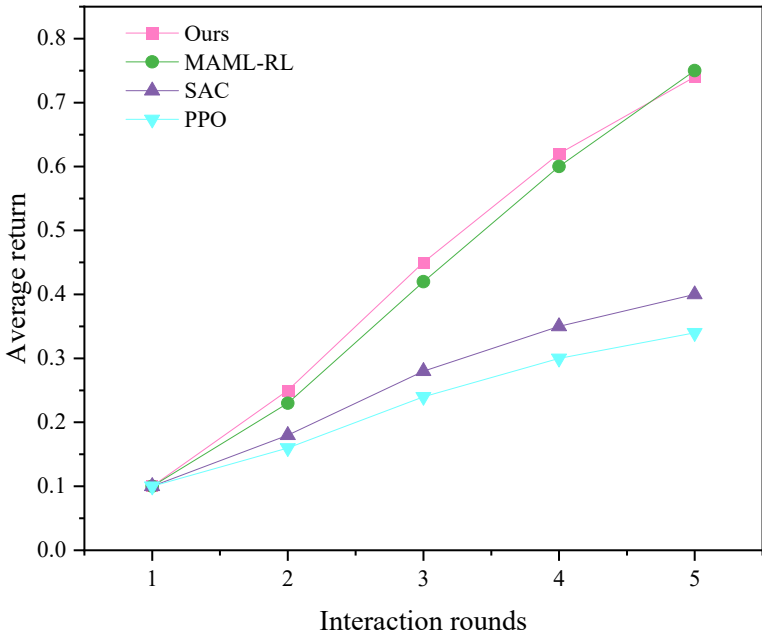
4.2.1 Ablation studies

To thoroughly analyse the contributions of each core component within the value-oriented meta-adaptive reinforcement learning framework proposed in this study, we conducted systematic ablation experiments. We constructed four key model variants for comparison: ‘no meta-learning’ (removing the meta-learning layer and training from scratch for each user), ‘no value-orientation’ (replacing the reward function with a single affect improvement reward), ‘single reward’ (using only effect rewards while removing engagement and cost rewards) and ‘fixed policy’ (employing a unified policy trained on the entire training set). Performance comparisons between these variants and the full model on the ESConv dataset are shown in Table 2.

Table 2 Ablation experiment results (mean $\pm$ standard deviation)

<i>Method variants</i>	<i>Emotional improvement</i> ($\uparrow$)	<i>User satisfaction</i> ($\uparrow$)	<i>Adaptation speed</i> ($\uparrow$)	<i>Safety score</i> ($\uparrow$)
Complete model	0.78	0.82	0.74	0.91
Zero-parameter learning	0.62	0.65	0.41	0.87
Value-free	0.71	0.74	0.69	0.73
Single reward	0.65	0.68	0.7	0.89
Fixed strategy	0.58	0.61	0.35	0.85

Figure 3 Learning curve of the metaprosthetic process (see online version for colours)

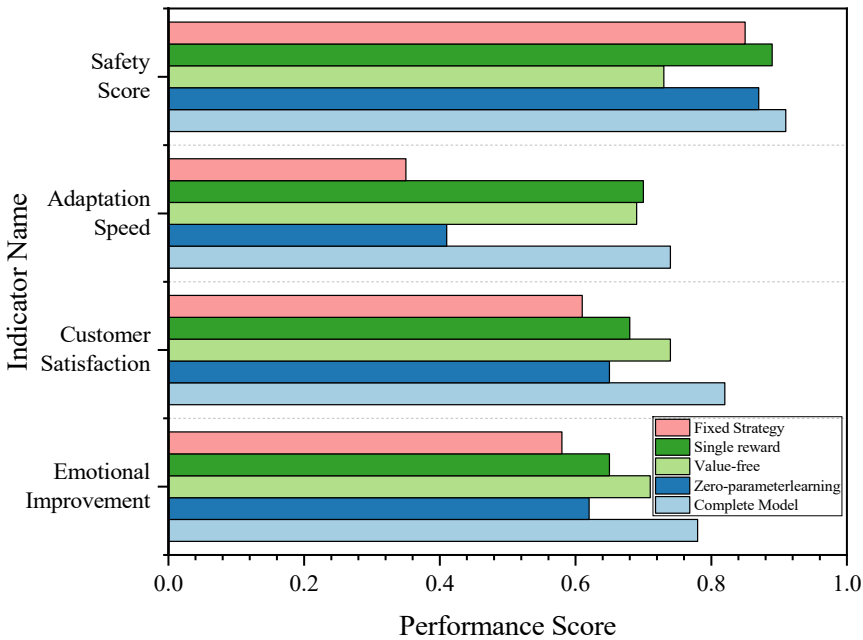


To more intuitively show the specific influence of each component on different evaluation performance, we visualise the data shown in Figure 2 in horizontal bar chart as shown in Figure 3. The horizontal bar chart is obvious to show the performance

deficiency of each ablation variant on different evaluation metrics, and the relative improvement of each component is more directly reflected.

As shown in Figure 4, the horizontal bar chart is visually used to show the important roles of each component. It is obvious that the bar length of ‘full model’ (dark blue bars) is the longest in three different metrics-‘emotional improvement’, ‘user satisfaction’ and ‘security score’, which is obviously better than other ablation variants. This directly validates the success of the whole integrated framework of meta-learning with value-oriented reward.

Figure 4 Performance comparison of components based on ablation experiments (see online version for colours)



Specifically, the ‘meta-learning-free’ variant had the largest performance drop in the ‘adaptation speed’ metric, and its bar length was only 55% of the full model. This result strongly shows that the meta-learning part is very important to achieve the fast personalisation of new users; after removing this part, the model regressed to a slow learning state and was almost like starting from scratch. The ‘value-free’ variant had the largest performance drop in the ‘safety score’, and its bar length was much shorter than other variants in this metric. This shows that the value-oriented reward function is an indispensable part of the intervention behaviour. Without the constraint of multi-objective safety reward, the model will adopt high-risk strategies to achieve more emotional improvements, and the model will lose in the safety score. Additionally, the ‘single reward’ variant (purple bar) also has obvious performance deficiencies in the ‘user satisfaction’ metric. This shows that if the model ignores user engagement and intervention cost even to some extent, even though it achieves some emotional improvements, it will still not be favoured by users in the long term and fail to get more positive feedback from users. Finally, the ‘fixed strategy’ variant’s performance is also

very poor, which shows that the extreme situation of personalisation is also very important in the field of emotional intervention.

5 Conclusions

This study addresses the core challenges of insufficient personalised adaptability and single-valued objectives in dynamic affective interventions by proposing a value-oriented meta-adaptive reinforcement learning framework. Experimental results demonstrate that on both the ESConv and DAIC-WOZ public datasets, this framework significantly outperforms all baseline models in two key metrics: emotional state improvement (0.78) and user satisfaction (0.82). It also achieves an outstanding safety score (0.91), validating its effectiveness and reliability in practical applications. Ablation experiments further confirm that the meta-learning mechanism is crucial for achieving rapid personalisation, with adaptation speed decreasing by approximately 45% when removed. Meanwhile, the value-oriented reward design is central to balancing intervention effectiveness with safety ethics; removing it causes safety scores to plummet by nearly 20%. These findings fully demonstrate the necessity and synergistic effects of each component within the framework.

In terms of theoretical contributions, this study achieves three major breakthroughs: first, it establishes a dual-level optimisation theoretical framework integrating meta-learning and reinforcement learning, providing a novel modelling paradigm for addressing individual variability in sequential decision-making tasks. Second, it designs a structured multi-objective reward function that systematically integrates intervention effectiveness, user engagement, safety assurance, and operational costs into the agent's optimisation goals, driving a paradigm shift in reinforcement learning from single-performance orientation to multi-value alignment. Third, it constructs a meta-adaptive learning algorithm tailored for psychological intervention scenarios. Through the synergistic optimisation of inner and outer loops, it achieves rapid capture of user dynamics and strategy adjustment with minimal interaction samples.

In terms of practical recommendations, this research provides clear guidance for developing next-generation intelligent psychological intervention systems. Healthcare institutions and digital health platforms can leverage this framework to build adaptive intervention systems that dynamically adjust support strategies by continuously learning individual users' response patterns, thereby enhancing intervention precision and user retention. Clinicians can also utilise this system as an auxiliary decision-making tool, optimising treatment plans by analysing the optimal intervention strategies recommended by the model. Furthermore, the framework's built-in safety constraints provide crucial reference for AI applications in high-risk domains, particularly in scenarios requiring stringent safeguards for user psychological well-being.

Acknowledgements

This work is supported by the Department of Education of Anhui Province 2023 Project under the 'Three-All Education' Comprehensive Reform and Ideological-Political Capacity Enhancement Program (No. sztsjh-2023-3-36).

Declarations

The author declares that she has no conflicts of interest.

References

- Amer, N.A., Shohieb, S.M., Eladrosy, W., Liu, S., Nam, Y. and Abdelrazek, S. (2024) ‘A gamified cognitive behavioral therapy for Arabs to reduce symptoms of depression and anxiety: a case study research’, *Digital Health*, Vol. 10, No. 4, pp.97–114.
- Arana-Catania, M., Sonee, A., Khan, A.M., Fatehi, K., Tang, Y., Jin, B., Soligo, A., Boyle, D., Calinescu, R. and Yadav, P. (2025) ‘Explainable reinforcement and causal learning for improving trust to 6G stakeholders’, *IEEE Open Journal of The Communications Society*, Vol. 6, No. 5, pp.4101–4125.
- Berkenkamp, F., Turchetta, M., Schoellig, A.P. and Krause, A. (2017) ‘Safe model-based reinforcement learning with stability guarantees’, *Physiology & Behavior*, Vol. 3, No. 14, p.374.
- Bertamini, G., Perzoli, S., Bentenuto, A., Furlanello, C., Chetouani, M., Cohen, D. and Venuti, P. (2025) ‘Child-therapist acoustic synchrony and response trajectories in autism intervention: an AI-based automated analysis using dynamic systems theory and affective computing’, *Emerging Trends in Drugs, Addictions, and Health*, Vol. 5, No. 12, p.591.
- Griefahn, A., Zalpour, C. and Luedtke, K. (2024) ‘Identifying the risk of exercises, recommended by an artificial intelligence for patients with musculoskeletal disorders’, *Scientific Reports*, Vol. 14, No. 1, p.521.
- Han, X., Zhao, Y., Gao, X., Wang, Y. and An, T. (2021) ‘The historical footprint and future challenges of water-energy-food nexus research: a bibliometric review towards sustainable development’, *Environmental Reviews*, Vol. 29, No. 1, p.719.
- Ige, T., Kiekintveld, C., Piplai, A., Wagler, A., Kolade, O. and Matti, B.H. (2024) ‘Towards an in-depth evaluation of the performance, suitability and plausibility of few-shot meta transfer learning on an unknown out-of-distribution cyber-attack detection’, *International Symposium on Networks, Computers and Communications*, Vol. 3, No. 17, p.719.
- Jeronimo, M.X.C., Basilio, A.Y.P. and Ruiz, J.A.Z. (2025) ‘Human talent management and the work performance of employees in a textile company in comas’, *Environmental Reviews*, Vol. 12, No. 15, pp.318–319.
- Ju, Z. (2024) ‘Effects of enhanced phytoremediation techniques on soil aggregate structure’, *Agriculture*, Vol. 14, No. 15, p.502.
- Khediri, N. (2017) ‘Towards an online emotional recognition system for intelligent tutoring environment’, *European Journal of Medical Research*, Vol. 6, No. 17, p.392.
- Kirilov, L., Guliashki, V. and Staykov, B. (2021) ‘Web-based decision support system for solving multiple-objective decision-making problems’, *Research Anthology on Decision Support Systems and Decision Management in Healthcare, Business, and Engineering*, Vol. 8, No. 2, p.317.
- Koussounadis, A., Redfern, O.C. and Jones, D.T. (2009) ‘Improving classification in protein structure databases using text mining’, *Bmc Bioinformatics*, Vol. 10, No. 23, p.618.
- Leafe, N., Pagnamenta, E., Taggart, L., Donnelly, M., Hassiotis, A. and Titterton, J. (2024) ‘What works, how and in which contexts when supporting parents to implement intensive speech and language therapy at home for children with speech sound disorder? A protocol for a realist review’, *The British Medical Journal-Open*, Vol. 14, No. 1, p.270.
- Liang, J. and Luo, B. (2012) ‘Toward a discourse shift in social gerontology: from successful aging to harmonious aging’, *Journal of Aging Studies*, Vol. 5, No. 6, p.295.

- Maqsood, Y., Usman, S.M., Alhussein, M., Aurangzeb, K., Khalid, S. and Zubair, M. (2024) 'Model agnostic meta-learning (MAML)-based ensemble model for accurate detection of wheat diseases using vision transformer and graph neural networks', *Computers, Materials & Continua*, Vol. 79, No. 5, pp.2795–2811.
- Meshram, R.A. and Alvi, A.S. (2024) 'Design of an iterative method for crop disease analysis incorporating graph attention with spatial-temporal learning and deep Q-networks', *International Journal of Intelligent Engineering & Systems*, Vol. 17, No. 3, p.719.
- Mostafa, H. and Lesser, V. (2011) 'A compact mathematical formulation for problems with structured agent interactions', *Digital Health*, Vol. 3, No. 12, p.518.
- Pitsch, H. (2024) 'The transition to sustainable combustion: hydrogen- and carbon-based future fuels and methods for dealing with their challenges', *Proceedings of the Combustion Institute*, Vol. 40, No. 1, p.36.
- Repp, A., Sheth, R.K., Szapudi, I. and Cai, Y.C. (2025) 'Indicator functions: distilling the information from gaussian random fields', *Journal of Aging Studies*, Vol. 9, No. 15, p.582.
- Shi, Y., Pei, H., Feng, L., Zhang, Y. and Yao, D. (2024) 'Towards fault tolerance in multi-agent reinforcement learning', *Scientific Reports*, Vol. 7, No. 14, p.293.
- Talace, N., Yekta, M.A. and Vaseghi, S. (2024) 'New insights into individual differences in response to chronic unpredictable mild stress in rats with respect to hippocampal BDNF and GSK β expression levels', *Physiology & Behavior*, Vol. 287, No. 2, p.179.
- Tassi, C. and Brouzos, A. (2025) 'Therapeutic factors in a group positive psychological intervention (ppi) for primary school students facing parental psychological economic hardship', *International Journal of Applied Positive Psychology*, Vol. 10, No. 3, p.392.
- Wu, J., Dong, Q., Gui, J., Zhang, J., Su, Y., Chen, K., Thompson, P.M., Caselli, R.J., Reiman, E.M. and Ye, J. (2021) 'Predicting brain amyloid using multivariate morphometry statistics, sparse coding, and correntropy: validation in 1,101 individuals from the ADNI and OASIS databases', *Frontiers in Neuroscience*, Vol. 15, No. 8, p.595.
- Yang, M., Tu, W., Qu, Q., Zhao, Z. and Zhu, J. (2018) 'Personalized response generation by dual-learning based domain adaptation', *Neural Networks*, Vol. 4, No. 31, p.926.
- Yelvington, M.L., Tyler, C., Jiale, H., Wood, R.E. and Stacey, R. (2024) 'The costs of caring: quantifying compassion fatigue and compassion satisfaction in burn therapists', *Journal of Burn Care & Research*, Vol. 12, No. 4, p.4.
- Yu, H., Zhao, W. and Zhao, Q. (2022) 'Distributed representation learning and intelligent retrieval of knowledge concepts for conceptual design', *Advanced Engineering Informatics*, Vol. 16, No. 8, p.519.
- Zhang, Y., Wang, Y., Yang, J., Li, Q., Zhou, M., Lu, J., Hu, Q. and Ma, F. (2025) 'Development and validation of machine learning-based risk prediction models for ICU-acquired weakness: a prospective cohort study', *European Journal of Medical Research*, Vol. 30, No. 1, p.382.