International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-025-00494-6

REVIEW

Reinforcement Learning in Education: A Systematic Literature Review

Anna Riedmann[1] · **Philipp Schaper**[1] · **Birgit Lugrin**[1]

Accepted: 16 June 2025

Abstract

We conducted a systematic literature review to examine the current research on the application of Reinforcement Learning (RL) in education. RL is a type of Machine Learning that trains an agent to take actions in an environment in order to maximize a reward signal. In recent years, researchers have explored the potential of RL for improving educational outcomes and developing personalized interventions. This systematic review (according to the PRISMA standard) surveys and evaluates 89 manuscripts from three databases (IEEE Xplore, Google Scholar, and ACM) published between 2000 and 2024 with predefined eligibility criteria. We examined the following objectives: (1) Educational contexts and evaluation strategies in RL-based educational applications, (2) impact and significance of RL-based applications for cognitive and affective variables, (3) RL algorithms and baselines used in the context of RL in education, (4) adaptation mechanisms in RL-based education, and (5) best practices for implementing RL in education. Our results suggest that RL has shown promise for a range of educational applications, such as enhancing learning outcomes or promoting student engagement. However, there are currently significant challenges and limitations to the use of RL in education, including methodological issues, and the need for broader and more large-scale deployments and evaluations with actual users relative to only using simulated data. Overall, this review provides a comprehensive overview of the current state of research on the application of RL in education and identifies areas where further research is needed to fully realize its potential as a tool for enhancing teaching and learning. Additionally, we present a set of best practices for the field, distilling key insights from our systematic review for practical application.

Keywords Reinforcement learning · Technology-supported learning · Systematic literature review · Education

✉ Anna Riedmann
anna.riedmann@uni-wuerzburg.de

[1] Socially Interactive Agents, Julius-Maximilians-Universität Würzburg, Würzburg, Germany

Published online: 10 July 2025

Introduction

Learning environments that adapt their content and/or feedback mechanisms to learners, are a common approach in the field of technology-supported learning. Over the last few years, there has been an increased interest in applying Reinforcement Learning (RL) in educational domains to optimize adaptive and personalized learning environments. RL is an area of Machine Learning (ML) and describes the aim of maximizing a reward signal by learning the optimal state-action-mapping in a trial-and-error process (Sutton & Barto, 2018). This method has been applied to improve adaptive learning environments with the overarching goal of enhancing the learning experience for learners, e.g., through adaptive difficulty and scheduling of learning content or personalized feedback. Therefore, RL offers a new approach to adaptive and personalized learning environments, with the potential to benefit learning outcomes and the overall learning process.

Through a systematic literature review, we provide comprehensive insight on the field of research by presenting both a descriptive overview of the application of RL in education, while offering an interdisciplinary perspective on potential application contexts and significance of reported results. We do not only examine the role of RL policies for instructional sequencing purposes, but also in terms of guidance provided, referring to problem solving steps taken by an RL agent, to aim for a more holistic approach in the field of RL in education. As a result of our systematic review, we synthesize a set of best practices for the field, summarizing the key insights from our systematic review to be applied in practice. Our review adheres to the PRISMA standard (Page et al., 2021), allowing for a standardized and systematic reporting of our literature review. PRISMA is a widely recognized framework for conducting systematic reviews, ensuring transparency, completeness, and accuracy in reporting results while supporting a reproducible review process. Further, PRISMA is linked to more comprehensive reporting of systematic reviews (Leclercq et al., 2019; Panic et al., 2013). Although concerns have been raised about an over-reliance on PRISMA guidelines (Teixeira da Silva & Daly, 2024), we believe that the benefits, such as comprehensive reporting and reproducibility, outweigh potential drawbacks (e.g., the risk of missing studies). To mitigate this, we carefully followed the guidelines outlined by Page et al. (2021).

The application of RL in education has been previously investigated by Doroudi et al. (2019), providing a thorough analysis of the use of RL for instructional sequencing, which refers to sequencing of learning content using RL to optimize the students' learning over the course of the learning environment. For an in-depth investigation of empirical studies comparing RL-based instructional policies to other sequencing methods, we direct readers to their work. Taking a different review approach, our study expands on this by broadening the scope and focusing on the latest advancements from the past five years to reflect the field's rapid development. We also refer the reader to Doroudi et al. (2019) for a historical perspective on RL. Concurrent to our review process, Fahad Mon et al. (2023), and Memarian and Doleck (2024) also investigated the utilization of RL

in education, which confirms the relevance of and ongoing activity in the research field. The authors provide descriptive surveys of the field of research, and we refer the reader to Fahad Mon et al. (2023) as well as Memarian and Doleck (2024) for a complementing overview with a focus on ethical concerns and general challenges. In addition, previous work from den Hengst et al. (2020) surveyed RL-based personalization approaches for a variety of domains, including education. Overall, while these reviews either cover a broader field, only address a specific section of applying RL to education, or set a different focus, the present work is set on the state of the art of integrating RL in education, considering both actions in the context of RL as instructional activities and as problem solving steps, while deriving guidelines for practitioners to facilitate the successful application of RL in educational practice. Coming from an interdisciplinary field, we additionally adopt the PRISMA guidelines for comprehensive reporting of results and reproducibility.

This systematic review is organized as follows. We briefly introduce RL in the context of Markov-Decision-Processes and present a short summary of associated methods and algorithms. Next, we present our systematic literature review approach and the corresponding criteria. In the scope of these criteria, we give an overview of application areas for RL in education, examine potential benefits and issues critically, and discuss results and implications.

Overview on Reinforcement Learning

Machine Learning generally can be subdivided into three main learning techniques, namely Supervised Learning, Unsupervised Learning and RL, as noted by Sutton and Barto (2018). The former refers to a concept of learning by training on labeled datasets, while Unsupervised Learning performs training based on independent information search in unlabeled data sets to uncover structure (Sutton & Barto, 2018). Akanksha et al. (2021) also describe Semi-Supervised Learning as a hybrid form of Supervised and Unsupervised Learning for application areas in which labeling is associated with high costs. RL can be considered as an additional type of ML where an agent uses an algorithm to decide for an action on the given environmental conditions and is able to learn from its past actions and experiences (Akanksha et al., 2021). As a result of taken actions, the agent receives a reward. In its aim of maximizing rewards in the long term, it typically converges to a policy π with the goal of approximating the optimal policy, determining the learning agent's optimal behavior at a given time, considering that this convergence is dependent on the (often limited) representation of states and actions (Iglesias et al., 2009a; Sutton & Barto, 2018).

RL is often considered in the framework of a Markov-Decision-Process (MDP). This includes an agent as a decision-maker selecting an action at each time step within an environment, thus resulting in different states of the environment and in the agent receiving rewards that it aims to maximize (Sutton & Barto, 2018). More specifically, MDPs are defined as a tuple $\langle S, A, R, p, \gamma \rangle$, with

S being a set of states in which the agent can be in, A being the set of actions it can select from in a given state at each time step t, and R being the numerical reward it receives in the next time step (Doroudi et al., 2019; Dulac-Arnold et al., 2021; Sutton & Barto, 2018). Variable p, with $p(s'|s, a)$, further denotes the probability distribution function of an environment when transitioning from one state s to another s' after taking an action a (Puterman, 2005; Sutton & Barto, 2018). A discounting factor $\gamma \in [0, 1]$ is applied for expected rewards (den Hengst et al., 2020; Sutton & Barto, 2018), resulting in the overall expected return G_t being denoted as $\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$ (Sutton & Barto, 2018). The interaction process in an environment can be described as the following: The starting point of the agent is a given state $s_0 \in S$ in an environment where the agent takes an action $a_t \in A$ at each time step t, resulting in the agent receiving a reward $r_t \in R$, and a state transition to $s_{t+1} \in S$ (François-Lavet et al., 2018); see Fig. 1 for an overview.

An important extension of MDPs can be defined by assuming that the policy must make decisions while having only partial information about the current state (e.g., the student's knowledge state is unknown); this is called a Partially Observable Markov-Decision-Process (POMDP; Papadimitriou & Tsitsiklis, 1987). A POMDP can be considered a "generalization of a [MDP] which permits uncertainty regarding the state of a Markov process and allows state information acquisition" (Monahan, 1982, p. 1). Thus, a POMDP is defined as a tuple $\langle S, \Omega, A, R, p_h, p_o, \gamma \rangle$, with A, R and γ being similar to MDP definition, S as the hidden state space and Ω being a set of observations (François-Lavet et al., 2018; Kaelbling et al., 1998). Further, p_h represents the transition probability between hidden states when an action is taken and p_o denotes the conditional probability of an observation occurring. While the agent observes and performs actions, it updates its internal belief state, representing the agent's probabilistic estimate of the true underlying state of the environment, given the history of actions and observations, i.e., the agent's previous experience (Kaelbling et al., 1998). Similar to the interaction process for a MDP, the agent starts in a given state $s_0 \in S$ in an environment where the agent initially makes an observation $\omega_0 \in \Omega$, occurring with probability $p_o(s_t, \omega_t)$, which is dependent on the state of the environment. Next, the agent also takes an action $a_t \in A$ at each time step t, resulting in the agent receiving a reward $r_t \in R$ discounted by $\gamma \in [0, 1]$, a state transition to $s_{t+1} \in S$, and the agent making another observation $\omega_{t+1} \in \Omega$ (François-Lavet et al., 2018). In MDP and POMDP, the agent aims to optimize the expected sum of discounted future rewards (Kaelbling et al., 1998).

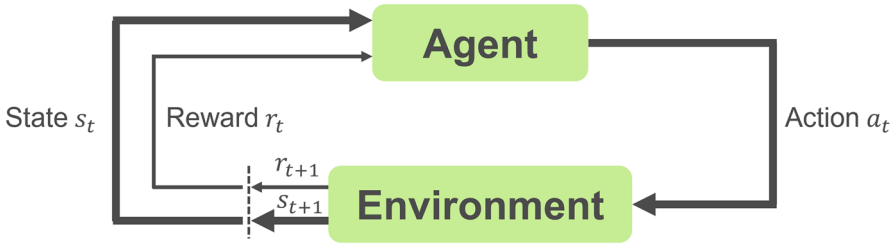


Fig. 1 RL Model Depicting the Agent-environment-interaction

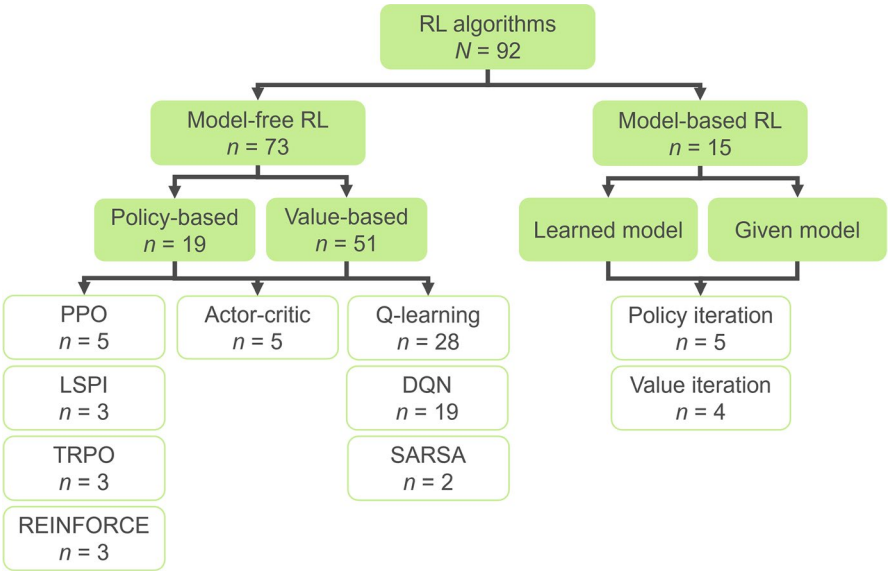
Methods for Solving MDPs

Methods for solving Markov-Decision-Processes can generally be distinguished as model-based (referring to methods of dynamic programming) and model-free (referring to Monte Carlo and temporal-difference learning methods), see Sutton and Barto (2018) for a more in-depth introduction. As the name suggests, model-based approaches use a model of the environment to predict rewards and state transitions (Akanksha et al., 2021), requiring a comparatively small sample size (Bennane, 2013; Lin et al., 2020). A model can either be given, provided that the reward function and probability distribution is known by the agent, or learned, requiring the agent to learn the model first by interacting with the environment before applying it during policy improvement (Zhang & Yu, 2020). Besides the sample efficiency (which is especially important when applying RL to real world problems), model-based approaches also provide several benefits, such as high safety and explainability as well as allowing for transfer learning (Moerland et al., 2023). However, model-based RL tends to lose asymptotic performance in the long run, especially for non-optimal models, and requires additional computational resources due to higher memory demands and an extensive number of tunable hyperparameters compared to model-free methods (Moerland et al., 2023). Apart from this, a major disadvantage is the lack of availability of an accurate model for many (real-life) environments (Zhang & Yu, 2020).

Model-free approaches do not infer the dynamics of the learning process using an environmental model but use interaction experience between agent and environment to estimate the optimal policy (Nguyen & La, 2019). Because model-free methods allow deriving the value function directly from interactions with the environment, they are applicable in contexts where no model of the environment is available, further entailing easy implementation and hyperparameter tuning (Nguyen & La, 2019). This results in several advantages, which is why this approach is used more often (Nguyen & La, 2019). However, it is notable that model-free methods experience challenges in terms of sample inefficiency, i.e., they require substantial interaction between the agent and its environment to learn valuable states (Afsar et al., 2023), which is often not feasible due to restrictions inherent to the application area (e.g., specific vulnerable target group, high costs of failure).

Reinforcement Learning Algorithms

The different methods covered in Sect. 2 encompass a variety of implementation techniques. To not exceed the scope of this paper, the following sections introduce the RL algorithms that we found to be most popular in the educational context as a result of our systematic literature review described in Sect. 3. Algorithms that were only mentioned in one publication are not discussed ($n = 15$). A taxonomy of the algorithms reported is shown in Fig. 2. We briefly elaborate on model-free as well as model-based methods. For a broader introduction to model-free and model-based methods, we refer the reader to Shakya et al. (2023), Sutton and Barto (2018), and Zhang and Yu (2020).



Note. Includes algorithms specified in publications included in this review, with three papers usingtwo different RL algorithms (Jung et al., 2024; Orsoni et al., 2023; Pogelt et al., 2024). One paper reported the use of both model-free and model-based methods, and two papers did not specify a method, resulting in N = 92. Only methods used in more than one paper are shown

Fig. 2 Overview of Types of RL Algorithms Used in More Than One Paper

Model-free methods comprise policy- and value-based methods as well as actor-critic methods. As part of policy-based methods, Trust Region Policy Optimization (TRPO) is a scalable algorithm for optimizing complex policies (Schulman et al., 2015). As noted by Schulman et al. (2015), this allows for "monotonic improvement, with little tuning of hyper parameters" (p. 1889). The algorithm iteratively optimizes a local estimate of the policy's expected return while applying a Kullback–Leibler divergence penalty (i.e., a penalty based on a measure of how much the updated policy deviates from the previous one) and has demonstrated consistent performance across several tasks in the field of robotic locomotion, such as learning simulated robotic swimming or hopping (Schulman et al., 2015). Building on this, Proximal Policy Optimization (PPO) uses clipped probability ratios, offering a conservative estimate (i.e., a lower bound) of the policy's performance (Schulman et al., 2017). The optimization process alternates between sampling data from the policy and conducting several optimization epochs on the collected data. PPO incorporates the benefits of TRPO, but offers a simpler implementation, broader applicability, and demonstrates improved sample efficiency (Schulman et al., 2017). Another model-free method is the REINFORCE algorithm, as an acronym for *REward Increment = Non-negative Factor × Offset Reinforcement × Characteristic Eligibility* introduced by Williams (1992). REINFORCE is a classic algorithm "whose update at time t involves just A_t, the one action actually taken at time t" (Sutton & Barto, 2018, p.

326). While REINFORCE allows for improved performance for small learning rates and guarantees convergence to a local optimum if the learning rate decreases over time, it can, however, have high variance, which leads to slower learning (Sutton & Barto, 2018). Further, Least-Squares Policy Iteration (LSPI) is a sample efficient and model-free method (Georgila et al., 2019), combining "value-function approximation with linear architectures and approximate policy iteration" (Lagoudakis & Parr, 2003, p. 1107). LSPI enables action selection without a model and facilitates incremental policy improvement, by effectively utilizing and reusing sample experiences collected through any method in each iteration (Lagoudakis & Parr, 2003).

Q-learning is a well-established and popular value-based algorithm (den Hengst et al., 2020), which aims to learn the optimal policy relying on previous interactions of the agent with the environment, thus maximizing the value of an action in a certain state (Shawky & Badawi, 2019). Q-values are computed to assess the effectiveness of an action a executed within a specific state s, with a Q-value being "the expected discounted reward for executing action a at state [s] and following policy π thereafter" (Watkins & Dayan, 1992, p. 56). As noted by Nguyen and La (2019), Q-learning methods are quite sample efficient and might therefore be suitable to be used in real-world applications. The State-Action-Reward-State-Action (SARSA) algorithm is an improvement of the Q-learning algorithm, which does not maximize the reward for the next stage of action to be performed but learns from the current set of actions in the current state and updates the Q-value for the corresponding states (Akanksha et al., 2021; Velusamy et al., 2013). Instead of mapping state-action pairs to a Q-value, the Deep-Q-Network (DQN) algorithm uses neural networks to map input states to action-Q-value pairs, with the ultimate goal of estimating the optimal action-value function (Mnih et al., 2015). The Bellman equation is used to iteratively update the network weights (Akanksha et al., 2021), see Bellman (1952) for an introduction. Another class of algorithms for MDP optimization are actor-critic methods, considered hybrid variants of actor-only and critic-only methods (Konda & Tsitsiklis, 1999). As noted by Murphy (2025), the term 'actor' denotes the policy component, while 'critic' refers to the value function. Actor-only methods aim to improve the policy directly based on rewards, while critic-only methods focus on estimating the value of actions or states for improving the policy indirectly. Actor-critic methods combine the advantages of actor-only and critic-only methods (Konda & Tsitsiklis, 1999), thus estimating both value and policy functions (den Hengst et al., 2020).

Model-based methods can either rely on a model that simulates the environment's behavior (or more broadly, one that enables predictions about how the environment will behave), or learn the model by interacting with the environment before applying it during policy improvement (Sutton & Barto, 2018; Zhang & Yu, 2020). Value Iteration and Policy Iteration are well-known model-based methods for solving MDPs. Value Iteration aims to find the optimal MDP policy and its value by directly iterating on the value function, and ultimately converges to the optimal values (Howard, 1960). It uses the Bellman optimality equation as an update rule (Sutton & Barto, 2018). Policy Iteration denotes the process of finding an optimal policy through a finite number of iterations (Sutton & Barto, 2018). Value Iteration is simpler than policy iteration, but may require more iterations, while Policy Iteration converges quickly, but requires solving the value function for each policy (Howard, 1960).

A Systematic Literature Review of the Integration of RL in Education: Method

The application of RL in education holds the potential to offer personalized and adaptable learning experiences, and there has been an increased interest for applying RL in this domain. As education evolves, RL stands out as a promising tool to enhance learning effectiveness and individualize the learning journey. We conducted a systematic literature review (SLR) to provide an overview of educational applications that apply RL methods, specifically addressing implemented RL techniques, type of RL-based adaptation and evaluation approaches. Thus, the objectives of this SLR are as follows:

1) Examine educational contexts and evaluation strategies in RL-based educational applications.
2) Analyze the impact and significance of RL-based educational applications for cognitive and affective variables.
3) Investigate RL algorithms and baselines used in the context of RL in education.
4) Classify adaptation mechanisms in RL-based education.
5) Derive best practices for implementing RL in education.

We refer to PRISMA (Page et al., 2021) as a standard for reporting SLRs, entailing an evidence-based minimum set of items for reviews evaluating the effect of interventions. PRISMA is a well-established framework for conducting systematic reviews, promoting transparency, completeness, and accuracy in reporting results while facilitating a reproducible review process. Additionally, PRISMA is associated with more thorough reporting of systematic reviews (Leclercq et al., 2019; Panic et al., 2013). While concerns have been raised regarding an over-reliance on PRISMA guidelines (Teixeira da Silva & Daly, 2024), we believe the advantages, such as comprehensive reporting and reproducibility, outweigh potential drawbacks (e.g., the risk of overlooking studies). To address these concerns, we adhered closely to the guidelines provided by Page et al. (2021).

We thus provide a systematic overview on RL-based applications in education, including work on both instructional sequencing as well as guidance-related approaches, such as selecting appropriate feedback or the type of activity. We also examine key challenges that can arise in real-world applications, such as in education, as identified by Dulac-Arnold et al. (2021). In particular, this includes (1) the application and training of RL methods in existing learning environments, often relying on limited amounts of interaction log data; (2) psychological processes, such as learning or motivation, are often only partially or not at all observable, thus require an estimation using existing data; (3) educational applications often pursue multiple optimization goals, such as improving learning, reducing time on task, while simultaneously reducing drop out; (4) RL policies need to react to the learner immediately when deployed in a real-world setting; and (5) learning environments implementing a RL model potentially affect the learners' experience and outcomes, thus requiring its actions to be easy to understand and interpret.

Eligibility Criteria

As determined by PRISMA (Page et al., 2021), we defined eligibility and exclusion criteria. Eligibility criteria determine what an article *must* include, while exclusion criteria specify what is *not allowed*. We formulated *four eligibility criteria* for studies to be included in our SLR: (1) Written in English and published in a peer-reviewed journal or conference proceeding; (2) with the field of application being in the area of education; (3) being addressed by implementing a RL algorithm, including all approaches that apply single-agent methods as defined in Sect. 2; and (4) the article had to be published between 2000 and 2024, in our aim to focus on the current research on applying RL to education (e.g., instructional sequencing) as well as the growing interest in using Deep Reinforcement Learning (DRL) for educational purposes in the field of RL in education (Doroudi et al., 2019). Additionally, *three exclusion criteria* were applied in our SLR. Articles were excluded due to (1) non-availability of full-text, (2) focusing on multi-agent systems, bandit algorithms, or student modeling as well as other ML algorithms exceeding the scope of this review, or (3) article status (work in progress, including roughly outlined frameworks or not yet implemented concepts). The manuscripts identified through our search were assessed for eligibility and exclusion criteria during both the screening process and the full-text review. Figure 3 provides an overview of the SLR process.

Information Sources and Search Strategy

We conducted a literature search in three databases: IEEE Xplore, Google Scholar, and ACM, thus covering a wide range of relevant research in the field. We ran the query on IEEE Xplore on 2nd of November 2021, on Google Scholar from 4th of November to 24th of November 2021, and on ACM from 25th of November to 30th of November 2021. To account for the latest trends in applying RL in education, we ran two additional queries. The second query was run on both IEEE Xplore and ACM on 19th of October 2023, and on Google Scholar from 19th of October to 9th of November 2023, incorporating only articles published from the beginning of 2021 to October 2023. The third query was run on both IEEE Xplore and ACM on 4th of March 2025, and on Google Scholar on 5th of March 2025, considering only papers published in 2023 and 2024. For all three runs at different stages during the review process of this paper, we used two combinations of keywords to search in the listed databases. These were *reinforcement learning AND education* and *reinforcement learning AND intelligent tutoring system*. Intelligent tutoring systems are commonly considered an overarching term for ML-driven adaptive learning systems (e.g., Graesser et al., 2012; Nwana, 1990) and we thus decided to include the term in our search strategy.

Study Selection and Exclusion

In a first step, we screened titles and abstracts of all articles found and compared them with our eligibility criteria, resulting in 184 papers considered for further

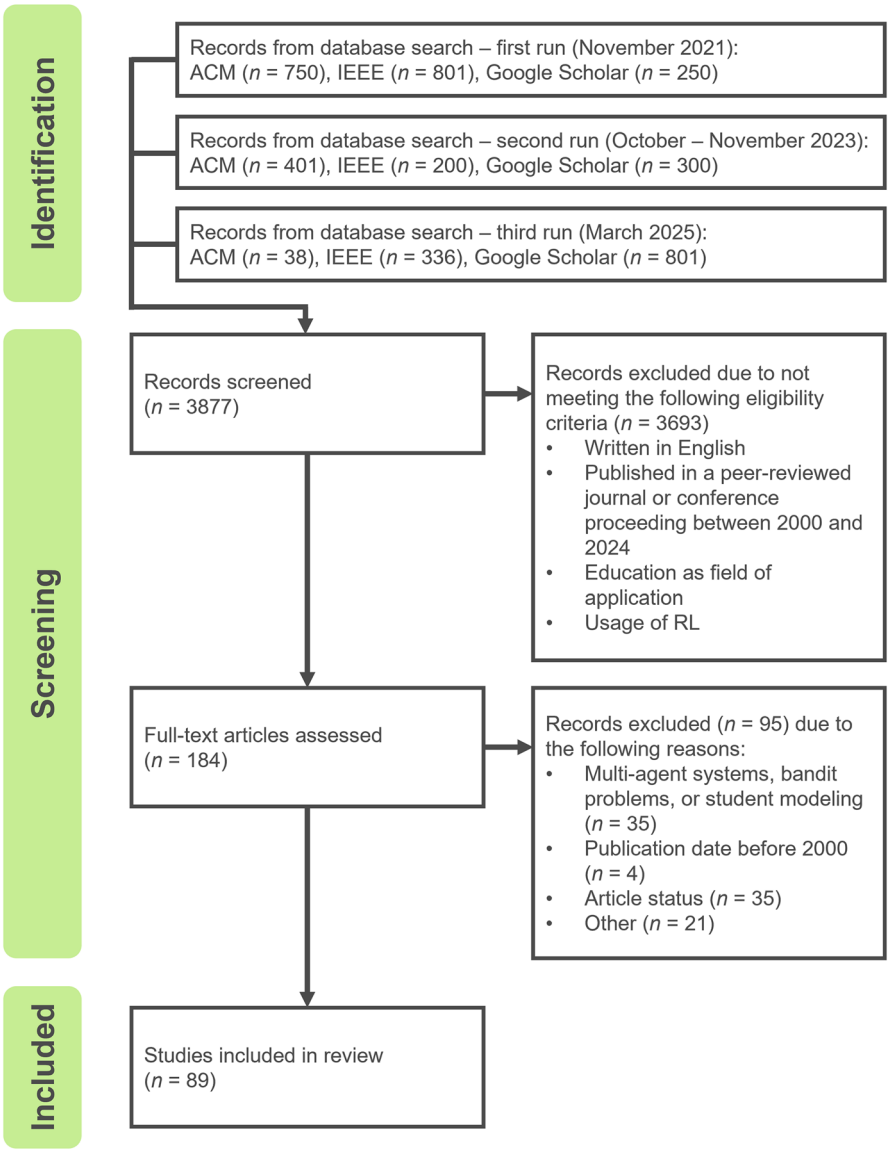


Fig. 3 Overview of the SLR Process

review, with duplicate articles already removed. Subsequently, the full texts of all selected papers were screened in regard to our exclusion criteria. This resulted in a final sample of 89 manuscripts for inclusion in this review. One author was responsible for selecting and screening the initial 184 articles. The final sample of 89 papers was screened by two authors. Disagreements were resolved by consensus through discussion.

Excluded studies comprised papers that contain *study proposals* or *conceptual work* (e.g., frameworks) (Amin et al., 2023; Barnes & Stamper, 2008; Boussakssou et al., 2020; Chi et al., 2008; Croy et al., 2008; Fenza et al., 2017; Gao et al., 2017; Geramifard et al., 2015; Jatzlau et al., 2019; Johnson & Zaiane, 2013; Koroveshi & Ktona, 2020; Kumar & Ahuja, 2020; Legaspi & Sison, 2002; Madani et al., 2020; Mirea & Preda, 2009; Mitchell et al., 2013; Mustapha et al., 2023; Ramachandran & Scassellati, 2014; Riedmann & Lugrin, 2023; Thede, 2002; VanLehn et al., 2007; Velusamy et al., 2013). We did also not consider *work in progress* (Condor & Pardos, 2022; Cui, 2023; Liu et al., 2022; Tang et al., 2022; Vijayan et al., 2018; Wang, 2018).

We excluded publications that focused on *student modeling*, i.e., estimating the student's knowledge state without optimizing instructional decisions, (Beck & Woolf, 2000; Dorça et al., 2013; Ferguson et al., 2009; Hostetter et al., 2023b; Kubotani et al., 2021; Lee & Brunskill, 2012; Li et al., 2023a; Perez et al., 2022; Sharma & Li, 2024), *multi-agent systems* exceeding our single-agent approach as defined in Sect. 2 (Beck et al., 2000; Bennane, 2013; Bittencourt et al., 2006; El Fazazi et al., 2021; El Fouki et al., 2017; Fahid et al., 2024; Hare & Tang, 2023; Hare et al., 2024; Kandel et al., 2022; Liu et al., 2023; Patel & Sajja, 2021; Priya et al., 2012; Zadem et al., 2023), or *Cellular Learning Automata* (CLA) (Minoofam et al., 2022). We further excluded papers addressing *bandit problems* (Belfer et al., 2022; Clement et al., 2015; Frenoy et al., 2016; Gao et al., 2018; Intayoad et al., 2020; Mazon et al., 2023; Mu et al., 2018, 2021; Roy et al., 2018; Schmucker et al., 2023; Soto Forero et al., 2024; Wang et al., 2020), such as multi-armed bandits and contextual bandits, as described by Sutton and Barto (2018). While bandit problems are often considered a simplified form of RL, conveying basic concepts of the field (e.g., exploration–exploitation tradeoff) (Sutton & Barto, 2018), we chose to exclude them from the present review as they lack state transitions and do not involve long-term planning (Lattimore & Szepesvári, 2020). Further, they appear to be not comprehensively represented by our systematic review approach, which we prioritized due to transparency and reproducibility.

Publications that did *not fit the educational context* of this review were also removed (Balaji et al., 2020; Bellotti et al., 2009; Frej et al., 2024; Haider et al., 2024; Maclellan & Gupta, 2021; Nie et al., 2023; Rojas-Barahona & Cerisara, 2014; Sarma & Ravindran, 2007; Scarlatos et al., 2024; Wang et al., 2018; Yang, 2024; Zhiyong et al., 2021). Papers applying *other Machine Learning concepts* rather than RL (Gao et al., 2023b; Kochmar et al., 2020; Leite et al., 2022; Loh et al., 2021; Nisansala & Morawaka, 2019; Yuh et al., 2024) were excluded as well as those *published before 2000* (Beck, 1997, 1998; Iglesias et al., 1995; Mishima & Asada, 1999). Manuscripts *without available full texts* were also excluded (Li et al., 2024a; Liu & Zoghi, 2023; Zhang, 2024). Finally, we did not account for publications being part of *already published work* (Abdelshiheed et al., 2023b; Chi et al., 2011; Iglesias et al., 2003; Shawky & Badawi, 2019; Tetreault & Litman, 2006b; Wang, 2014a; Zhou et al., 2021) or *unpublished* papers (Mandel, 2017).

Scope of the Review (Data Items)

The scope of our review spans six main outcome domains: The educational context, evaluation strategy, the aim of the study represented by the considered psychological and technical concepts as well as the reported results, the type of algorithm, and the type of adaptation. In terms of educational context, we assessed the *subject of the learning environment* and the *educational target group*. We clustered the reported target groups in preschool, elementary school, middle school, and high school education, as well as college and high education, and adult education. Further, we examined the *evaluation strategy* applied in the respective study. Specifically, we grouped the papers depending on whether the evaluation was performed (1) "live", i.e., interacting with actual learners in the real world, (2) based on real user data, or (3) using a simulator of user behavior, referring to the categorization proposed by den Hengst et al. (2020).

We also noted the *aim of the study* (if explicitly stated by the authors), such as benefitting motivation, increasing learning gain or improving the performance of the algorithm, allowing for an overview of the most researched psychological and technical concepts. In terms of *reported results*, we initially categorized the selected articles in five aspects, based on the review of Doroudi et al. (2019). That comprises studies with (1) at least one RL policy statistically significantly outperforming all baseline policies, (2) no significant differences, but a significant aptitude-treatment (ATI) effect (i.e., RL policy significantly outperforming baselines for low performing students), (3) mixed results with RL policy outperforming some, but not all baselines, (4) no significant differences, or (5) baseline(s) outperforming the RL policy. Because a high number of papers included in the review did not include any statistical analysis, we additionally added the category (6) no statistical comparison to differentiate between papers that did not find a significant effect (as a result of statistical tests) and those that did not perform or report any calculation at all, resulting in a total of six different categories. We also recorded effect sizes (specifically Cohen's d) and the respective 95% confidence intervals, either provided in the paper or calculated from descriptive values reported in the particular paper using the online calculator of Lenhard and Lenhard (2017). We initially planned to conduct synthesis methods to statistically evaluate the impact of RL on education through all publications, however, we refrained from it due to the following reasons: (1) The majority of papers employed several baselines, investigated multiple dependent variables, and reported varying numbers of outcomes of interest, deeming univariate meta-analysis impractical (Riley, 2009), (2) over half of the papers did not conduct statistical tests and/or report descriptive values, such as means, standard deviation or correlation, as well as (3) the non-availability of within-study co-variances which are prerequisite for performing multi-variate meta-analysis (Mavridis & Salanti, 2013).

We assessed articles based on the *type of RL algorithm used* (exclusively or additionally to other ML concepts) and whether it is model-based or model-free. Generally, we investigated whether researchers used RL methods or additionally combined

them with Deep Learning. We also included the *baseline(s)* used to compare with the proposed RL method. We categorized baselines reported in the papers in random baselines, baselines defined by domain experts (as denoted in the respective manuscript), other RL approaches, baselines that involved no adaptation (e.g., traditional lecturing or fixed curricula), heuristic baselines, or other types of baselines, such as other ML methods.

We further focused on the *type of adaptation* presented in the paper, subdividing it into two main categories: Content-related and guidance-related. We base our classification on Doroudi et al. (2019), who separated papers by the kind of actions taken in an educational context. With Doroudi et al. (2019) concentrating on research work that considers actions as instructional activities selected by a RL agent, we extended the scope and also included articles defining actions as problem solving steps to allow for a more holistic approach on RL in education. This categorization is also supported by Spain et al. (2021), distinguishing between systems that "present learners with different types of instructional feedback, hints, and faded worked examples […] [and] new learning content that is tailored to a learner's current skill" (p. 2). Singla et al. (2021) apply a similar categorization by also listing the personalization of the curriculum (i.e., instructional sequencing) and the provision of hints and feedback as main research directions when applying RL in education. Further, they extend our classification by student modeling (which we excluded for this paper as it exceeds the scope of our review, see Sect. 3.3), generating learning content (which does not fit the aim of our review), and A/B testing. The latter refers to the experimental comparison of different educational interventions, which employ adaptation through bandit algorithms, and is thus excluded from this review. In the scope of our review, content-related papers thus comprise studies that address instructional sequencing (also referred to as reinforcement scheduling, as noted by Bassen et al. (2020)) of learning activities to create personalized learning plans for students. These plans can include selecting the most appropriate content and adjusting the pace of learning and/or difficulty based on the student's performance. Thus, the content-related category focuses on the usage of RL to adapt the learning content to be presented. We further predefine the guidance-related category to include papers that employ RL to induce pedagogical strategies (e.g., hint generation or behavior adaptation, personalized feedback and guidance), as well as papers focusing on RL-based tutorial planning, defined as "how instructional feedback and support are structured and delivered to learners" (Spain et al., 2021, p. 2). In summary, the content-related category focuses on papers leveraging RL to adapt the learning content itself, whereas the guidance-related dimension involves papers on RL-based pedagogical strategies, such as selecting the type of activity or hint to provide in response to a given learning task or problem.

As additional variables, we collected data on *general paper information*: Author, year, and source of publication.

Results

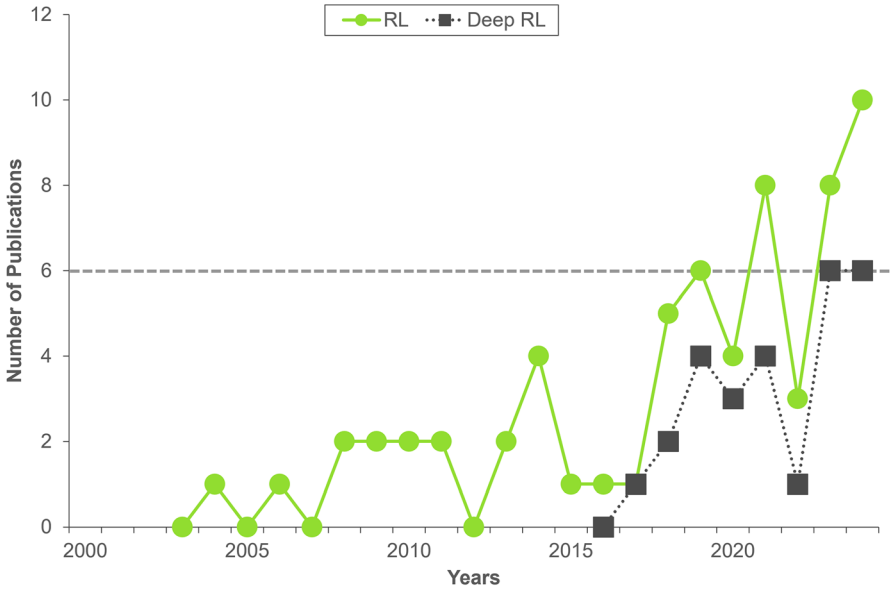
We group our results in scope of the review criteria, which corresponds to the outcome domains as noted in PRISMA (Page et al., 2021). The publications considered for this review were published over a period of more than 20 years, including six papers from 2000 to 2010 (excluding 2010), 11 papers between 2010 and 2015, and the majority of reviewed publications ($n = 72$) from 2016 to 2024. This results in a total of 89 papers included in this review. Appendix B contains a comprehensive overview of all included articles organized according to the specified data items for this review.

Focusing on the usage of Deep Learning when applying RL, we explored whether researchers employed RL techniques either individually or in conjunction with Deep Learning. As noted by Doroudi et al. (2019), there seems to be a trend in the use of RL for education (specifically when scheduling learning content) with researchers applying DRL to RL problems. This seems to be the case for the papers included in this review, with DRL demonstrating a slight increase in usage over the years. However, classic RL appears to be applicable for various educational scenarios, with a continuous upward trend until 2024, see Fig. 4 for a progression over the last 20 years.

Educational Context and Evaluation Strategy

RL methods are applied in different learning environments tackling a variety of learning subjects. Most popular learning topics are math ($n = 27$) and language learning ($n = 11$). We also noted physics ($n = 3$), biology and geology ($n = 3$), and computer-related topics ($n = 6$) in more than one publication. Thus, nearly half of all papers that specified a learning subject ($n = 62$) target STEM education ($n = 39$), with STEM denoting subjects related to science, technology, engineering, and mathematics. The remaining papers, which have defined a clear field of application, address different topics ($n = 12$), such as calligraphy training or basic counseling skills, while 27 papers did not specify a learning subject.

These RL integrated learning environments address a variety of different target groups, ranging from very young learners, such as preschool ($n = 1$) and elementary school children ($n = 6$), to adults ($n = 8$). Most learning environments target college or high education learners ($n = 30$), while there exist only few RL powered learning resources for high school ($n = 3$) and middle school ($n = 5$) education. However, nearly half of the reviewed publications did not specify an educational target group ($n = 36$). Table 1 displays an overview of the target group distribution across the different learning environment domains.



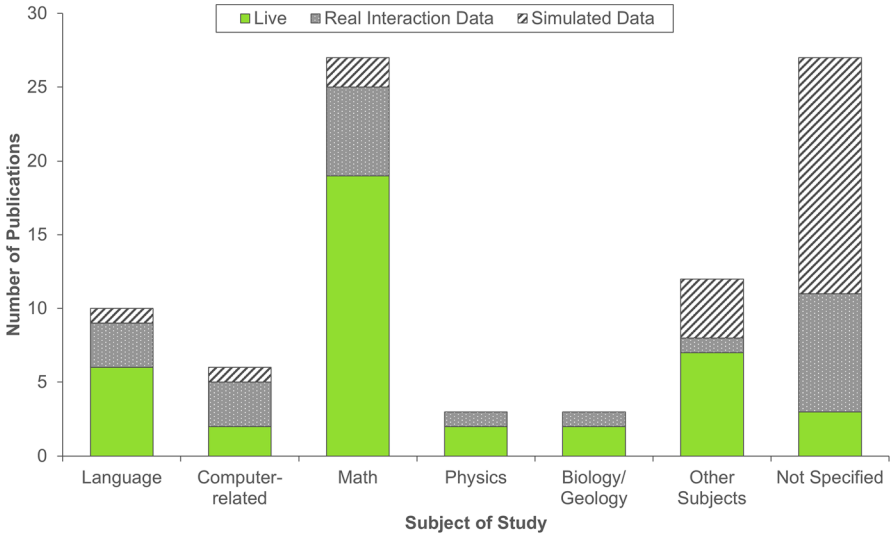
Note. One paper reported the use of both RL and DRL methods, resulting in $N = 90$

Fig. 4 Application of RL and DRL Methods in an Educational Context Distributed Over Years

With regard to the evaluation strategy realized in the reviewed papers, we classified whether evaluation of the RL policy was performed while interacting with learners in the real world (i.e., "live", $n = 41$), based on real interaction datasets ($n = 23$), or with simulated users ($n = 24$). One paper did not specify their evaluation approach. Figure 5 illustrates the distribution of the evaluation approaches across the learning environment subjects.

Table 1 Distribution of the Educational Target Groups Across Learning Subjects

	Language	Computer-related	Math	Physics	Biology/geology	Other subjects	Not specified	Total
Adults	3	1	-	-	1	2	1	8
Higher education	1	1	18	3	-	5	2	30
High school	-	-	2	-	-	1	-	3
Middle school	1	-	2	-	2	-	-	5
Elementary school	3	-	2	-	-	1	-	6
Preschool	1	-	-	-	-	-	-	1
Not specified	2	4	3	-	-	3	24	36
Total	11	6	27	3	3	12	27	89



Note. One paper did not specify the evaluation approach, resulting in $N = 88$

Fig. 5 Publications Sorted by Evaluation Strategy ("Live", Real Interaction Data, Simulated Data) Distributed Across the Learning Environment Subjects

Considered Concepts and Level of Significance

Several papers included in this review aimed to investigate the impact of RL on one or more concepts, comprising learning or performance, engagement and/or motivation or overall quality of interaction, learning efficiency (in terms of effectively accomplishing learning goals with low effort and time), time on task, or task completion. It should be noted that papers could be considered for more than one category because they explore multiple concepts. The majority of publications ($n = 51$) examined the impact of RL on a learning-related outcome variable, such as learning gain or performance in the learning environment (e.g., score). Learning efficiency was researched 12 times and affective variables, such as engagement, motivation, and/or quality of interaction, in 11 papers. Some publications also investigated time on task ($n = 12$) and the task completion rate ($n = 4$). Thirty-one studies assessed the performance of the RL algorithm used.

In terms of level of significance, over half of all included papers ($n = 54$) did not statistically analyze their results for significance, specifying only descriptive values (e.g., means and standard deviations) or figures, displaying different policies with the proposed policy visually outperforming the baseline(s). Please note that our review and the resulting findings are based on the information we have been able to extract from the papers. We have not conducted or recalculated any statistical tests to test for significance. Of the remaining 35 papers testing for statistical significance, we found 18 publications with at least one RL policy significantly outperforming all

baselines and two studies demonstrating a significant ATI effect (RL policy benefits low-performing students). Eleven papers reported mixed results and only four did not find a significant effect using statistical comparisons to test for significance.

Clustering papers that tested for statistical significance by the learner-related concepts investigated (e.g., learning gain, motivation), RL approaches appeared to be comparatively most effective for addressing affective variables, with 63% of papers reporting statistical significance, followed by task completion (50%), learning and performance (48%), and time on task (44%). Figure 10 provides an overview of the degree of significance of different concepts, clustered in the type of adaptation.

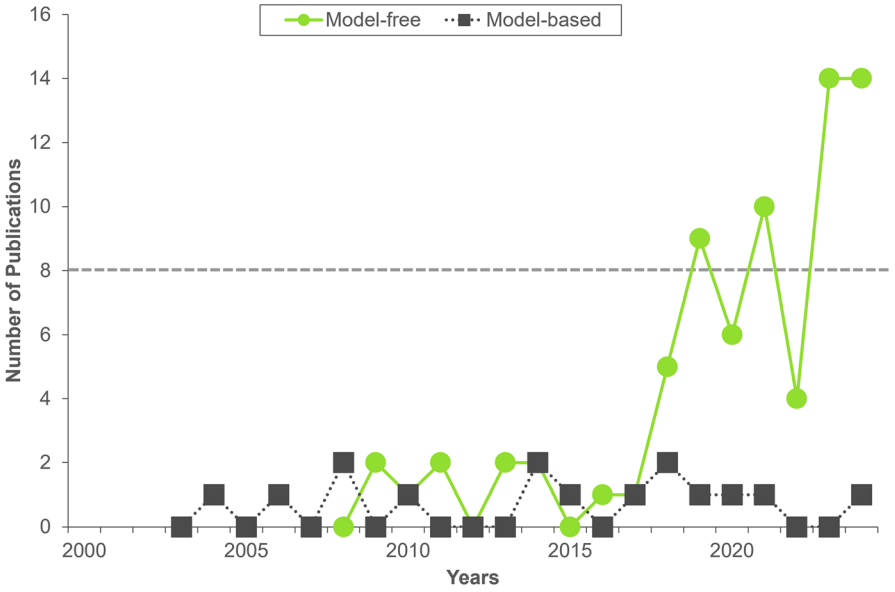
Regarding the reward functions used in the papers included in our review, we assessed the reward functions' elements of papers that reported RL approaches statistically outperforming one or more baselines ($n=18$). Learning gain, particularly normalized learning gain, which quantifies students' progress regardless of their initial competence, appeared to be an effective reward source and was used in over half of the papers ($n=10$). One approach additionally applied a penalty based on the number of materials assigned. Furthermore, accuracy, often combined with other factors like time or consecutive incorrect attempts, also proved effective, being part of the reward function in three approaches. Two RL approaches defined rewards based on both learning gain and student engagement, while one focused solely on engagement. Two papers incorporated other elements into their reward functions, such as students' preferred (feedback) template style.

Type of Algorithm and Baselines

The majority of papers included in this review followed a model-free approach ($n=72$), while 14 publications reported the use of model-based methods. One paper used both model-free and model-based approaches, and two did not specify the technical approach. Figure 6 depicts the progression of the use of these methods over time, illustrating a particularly prominent increase of model-free methods over the last six years ($n=61$). A major number of studies implemented value-based approaches ($n=51$), with Q-learning being the most frequently mentioned algorithm ($n=28$). A taxonomy of most frequently reported algorithms can be seen in Fig. 2.

Focusing only on papers that tested for statistical significance, 24 used model-free approaches and nine publications applied model-based methods, one paper used both model-free and model-based approaches, and one did not specify the technical approach. Figure 7 displays the (percentage) distribution of the results according to the respective application method.

We further examined the baselines used to compare with the suggested RL policy in all reviewed papers. While half of the reviewed papers ($n=46$) named one baseline, 25 RL policies were compared to two or more baselines. Eighteen publications did not specify any baseline. Overall, random baselines were used by one-third of all reviewed papers ($n=31$), expert-crafted baselines were used 19 times, and RL induced policies were compared to other RL approaches 20 times. Some papers also used heuristic policies ($n=4$) or other baselines ($n=13$), and 14 publications compared their RL-based



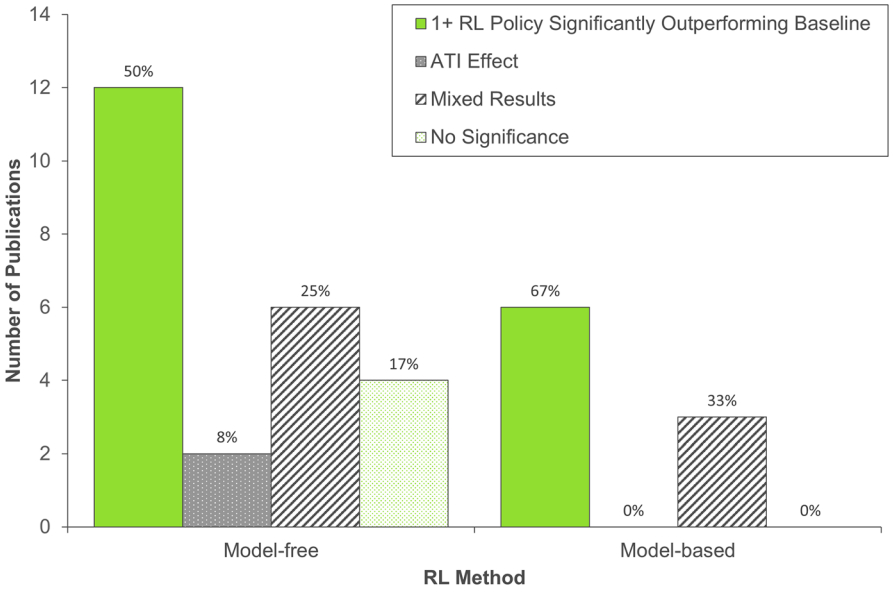
Note. One paper did not specify the evaluation approach, resulting in $N = 88$

Fig. 6 The Use of RL Methods Over Time

condition to a non-adaptive control group. Figure 8 provides an overview on the type of baselines used or combinations thereof, distributed according to the respective evaluation strategy used, including only publications that found a significant effect for one or more RL policies ($n = 18$).

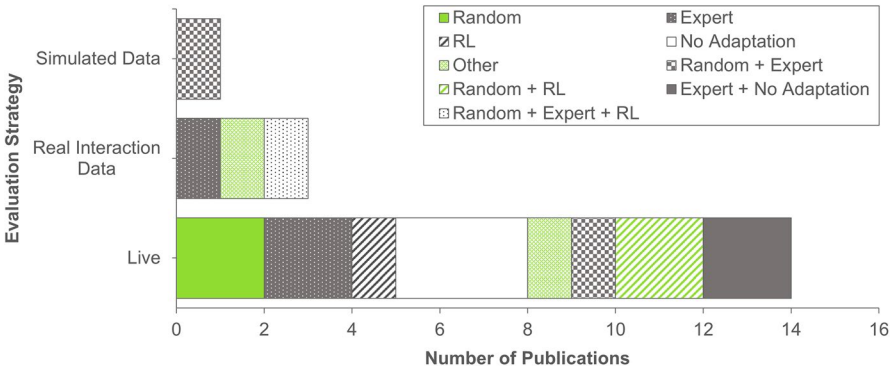
Among the studies that were accompanied by statistical testing ($n = 35$), RL approaches outperformed 57% of other RL policies, half of the expert and non-adaptive baselines (50% each), 47% of random baselines, and all other approaches used as a baseline. It should be noted that several papers employed more than one baseline.

For a more specific overview of the impact of RL on learning-related outcome variables compared to baseline policies, we additionally assessed all papers on given effect size(s) for statistical comparisons (if conducted) and respective 95% confidence interval. This excluded publications that did not conduct either statistical analysis in general, compared other outcome variables rather than post-test performance, or provided insufficient descriptive data (mean, standard deviation, sample size) to perform calculation of effect size. Publications that did not report on "live" evaluation results were excluded. Cohen's d was chosen as it is a popular effect size measure and was used in the plurality of included publications that reported effect sizes. Referring to Cohen (1988), the effect size d can be classified from small effects ($d = 0.2$) over intermediate ($d = 0.5$) to large effects ($d = 0.8$), according to the magnitude of the respective Cohen's d reported.



Note. Percentages indicate the relative frequency of the results using the respective RL method. Papers that did not test for statistical significance ($n = 54$) or specify the technical approach ($n = 1$) were omitted in this analysis as well as papers that used both model-free and model-based methods ($n = 1$)

Fig. 7 Results Clustered Based on Type of RL Method



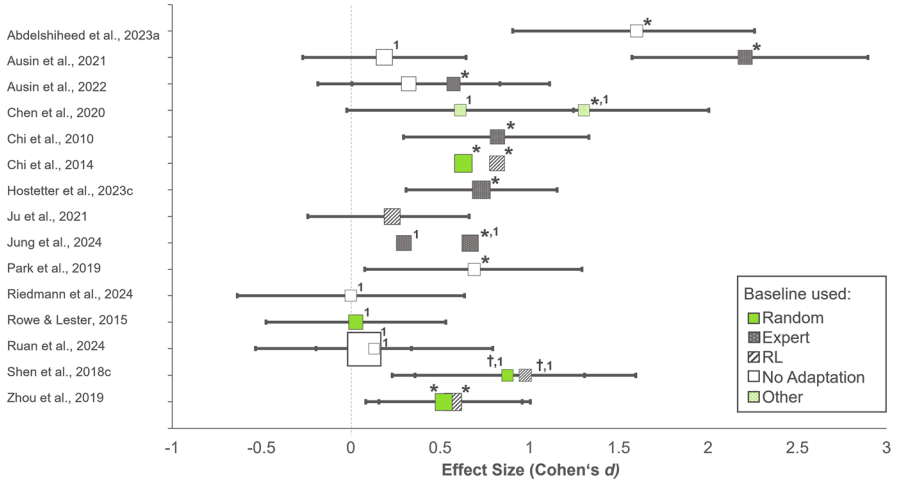
Note. Total number of papers reporting significant differences is 18

Fig. 8 Number and Type of Baselines (or Combinations Thereof) that Were Significantly Outperformed by a RL Induced Policy Distributed Across the Evaluation Strategy Used

We identified 15 papers to be suitable for our overview of effect sizes (see Fig. 9). It is notable that a majority of publications finding significant differences reported at least one intermediate ($n=5$) or large effect ($n=6$). It should further be noted that multiple effect sizes per paper may arise from having multiple baselines. For seven papers, effect sizes were calculated from descriptive values reported in the particular paper using the online calculator of Lenhard and Lenhard (2017).

Type of Adaptation

Considering the type of adaptation that the RL integrated environments described in the reviewed publications aimed for, we separated in (1) guidance-related adaptation, comprising papers that focus on guidance-related adaptation to the learner, for example through hints, or type of feedback or activity selected, and (2) a content-related category including all studies that investigated the RL induced scheduling of learning content. Our results indicate that there is a higher amount of papers using RL to adapt learning content to a specific group of learners. We identified 53 papers as content-based and 36 publications as guidance-related. Further detailed information about the papers clustered in level of significance and type of adaptation can be found in Table 2. See Fig. 10 for an overview of the degree of significance of different concepts, clustered in the type of adaptation.

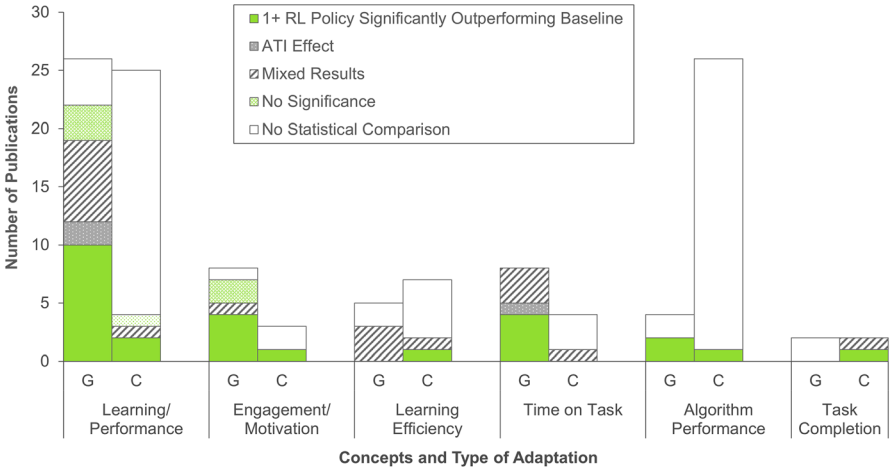


Note. Size of data points represents respective sample size. Pattern and/or color of data points indicates the type of baseline compared against. Includes only papers that conducted a "live" evaluation. * Indicates significance for $p < .05$. † Indicates significant ATI effect. 1 Cohen's d calculated by authors

Fig. 9 Cohen's d Effect Sizes and Respective 95% Confidence Intervals for Post-test Score Comparisons on a Learning-related Outcome Variable

Table 2 Papers Clustered in Level of Significance and Type of Adaptation

	Publications	Guidance-related	Content-related	Total
1 + RL policy outperforming baseline	Abdelshiheed et al. (2023a), Bassen et al. (2020), Chen et al. (2020), Chi et al., (2010, 2014), Efremov et al. (2020), Gao et al. (2023a), Gkatzia et al. (2013), Hostetter et al., (2023a, 2023c), Ju et al. (2021), Mandel et al. (2014), Park et al. (2019), Ravari et al. (2021), Sawyer et al. (2017), Wang (2014b), Yessad (2023), Zhou et al., (2019, 2020)	14	4	18
ATI effect	Shen et al., (2018a, 2018c)	2	0	2
Mixed results	Alam et al. (2024), Ausin et al., (2019, 2021, 2022), Chakraborty et al. (2021), Gordon et al. (2016), Jung et al. (2024), Rowe and Lester (2015), Shen et al. (2018b), Whitehill and Movellan (2018), Zhang and Goh (2021)	8	3	11
No significance	Georgila et al. (2019), Oralbayeva et al. (2022), Riedmann et al. (2024), Ruan et al. (2024)	3	1	4
No statistical comparison	Ai et al. (2019), Barnes et al. (2008), Caro et al. (2023), Efremov et al. (2020), Fernandes et al. (2023), Flores et al. (2019), Fotopoulou et al. (2020), Gong et al. (2023), Han et al. (2020), Huang et al. (2019), Iglesias et al., (2009a, 2009b), Intayoad et al. (2018), Islam et al. (2021), Jeewantha et al. (2021), Kakdas et al. (2024), Kim et al. (2021), Koroveshi and Ktona (2021), Li et al. (2023b), Li et al. (2024b), Liang and You (2024), Liu et al. (2018), Liu et al. (2019), Malpani et al. (2011), Martin and Arroyo (2004), Ming and Hua (2010), Mohana et al. (2024), Niu and Cao (2022), Orsoni et al. (2023), Oyuga Anne and Maina (2021), Pan and Yang (2024), Pérez et al. (2024), Pietquin et al. (2011), Pögelt et al. (2024), Pu et al. (2020), Raghuveer et al. (2014), Reddy et al. (2017), Shawky and Badawi (2018), Shen et al. (2021), Shin and Bulut (2022), Spain et al. (2021), Stamper et al. (2008), Su et al. (2013), Sun et al. (2024), Tang et al. (2019), Tetreault and Litman (2006a), Vassoyan et al. (2023), Wan et al. (2023), Wang et al. (2023), Wang and Song (2024), Wu et al. (2024), Yantao and Wei (2024), Zhang (2023), Zhang and Goh (2019)	9	45	54



Note. Results labelled as 'G' describe the guidance-related category, while 'C' comprises all publications with content-related adaptations. Many papers examine more than one concept, thus multiple counting is possible

Fig. 10 Significance of Different Concepts Clustered in the Type of Adaptation

Discussion

In our systematic literature review, we present an overview of the application of RL methods in educational contexts. To this end, we reviewed 89 publications matching our eligibility criteria and analyzed them according to the following factors: Significance of results in the scope of the educational context and evaluation strategy, type of algorithm, considered concepts, and type of adaptation.

Educational Context and Evaluation Strategy

We note a general increase of studies that use RL in the educational domain over time. While RL methods have been applied in a variety of educational fields, STEM-related topics, such as math, physics or biology, seem to be the most popular learning subjects for RL-based educational applications. STEM education is often focused on teaching segregated theoretical concepts, conveyed through direct instruction (Nadelson & Seifert, 2017). Many current educational systems continue to rely on outdated instructional approaches, often failing to capture students' interest and engagement, and can thus considerably benefit from a technology-supported approach, allowing for application and experiential learning (Vahidy, 2019). The majority of STEM-related applications also involve mathematical tasks, enabling easy automatic generation of tasks with simultaneous estimation and adjustment of difficulty. This probably explains the high number of STEM-related learning environments in papers included in this review. While

these results indicate the feasibility of RL-based learning environments, they also highlight the need for further research on the impact of RL in other knowledge domains. Future work might therefore focus on extending and enhancing existing approaches as well as investigating the transferability of concepts for novel applications of RL in education beyond the STEM umbrella.

Most RL-based educational applications, addressing a specific group of learners, target adults or higher education students. RL often requires a great amount of data for training, requiring applications that can easily record learner data in large quantities. This is the case in Massive Open Online Courses (MOOCs) that often target adult learners (for an overview, see Li (2019) and Ruipérez-Valiente et al. (2022)), allowing to capture massive amounts of interaction data of real students, thus facilitating the implementation and use of RL methods. While MOOCs are useful to easily generate data, in return, they are associated with very heterogeneous participants and many dropouts. Exploring other contexts, e.g., schools, might yield more homogeneous yet smaller groups. However, learning starts at a young age, requiring educational environments to adapt to (young) children's affective state and learning progress, as noted by Guran et al. (2021). Future work applying RL should therefore expand the focus on applications targeting children to support their overall learning process.

Nearly half of the reviewed publications did not target a specific group of learners and 30% did not specify a learning environment topic. This might be due to the large number of highly technical publications proposing minor variations of existing algorithms to be used in an educational context, while also evaluating their performance with simulated students, which does not require elaborating on the educational context or the respective target group. However, it is notable, that using RL methods in the domain of education offers unique challenges that might not be easily solved by applying common RL approaches. As summarized by Singla et al. (2021), this comprises the lack of simulations which adequately reflect the respective learning context to efficiently train RL models, large and continuous state spaces that are not always completely observable (e.g., knowledge or motivation), as well as methodological and also ethical concerns.

Half of the reviewed publications evaluated their approaches in a "live" scenario, i.e., with learners in the real world. While simulation and interaction datasets allow training and evaluating RL models risk-free and in a resource-efficient way, it is essential to also evaluate these methods in a real-world context, which is why this development is promising. As noted by Dulac-Arnold et al. (2021), real-world RL systems are often difficult to simulate, lacking concrete rules and conditions, which artificially limit the application environment in a real-world context. In the scope of the challenges identified by Dulac-Arnold et al. (2021) that apply to the use of RL in education (see Sect. 3), we observed that educational applications presented in the reviewed publications often pursued multiple optimization goals, for example improving learning while reducing time on task, and although various measurement methods have been applied, not all aspects of user interaction, such as psychological processes, are always observable. Considering that the application of RL methods in an educational context aims to benefit the performance and affective state of learners, it is necessary for real learners to actually use such a learning system,

thus requiring test runs and pilot studies to optimize the applied RL approach for such use cases. Overarching the goal of improving the learning process of real-world learners, the increasing amount of real-world evaluations is highly desirable.

Type of Algorithm

With a view to RL methods applied in all reviewed papers, a majority used model-free methods, which seem to have experienced a massive increase in popularity for educational purposes over the last six years. The increased usage of model-free RL is also evident in other application areas, such as social robotics (Akalin & Loutfi, 2021). This development might be due to the limited availability of environmental models that can be used in educational contexts. While model-based RL benefits sample efficiency, there is often no suitable model available or it is too difficult to learn (Nguyen & La, 2019), especially when considering particular target groups such as young children where it is not feasible to devise an adequate model to map a child's learning process without any prior interaction (Chen et al., 2020). Thus, model-free methods are applied more often, because they are able to derive the value function directly from interactions with the environment and allow for easy implementation and tuning of hyperparameters.

However, model-based approaches can bring additional benefits in the educational context, such as requiring less interaction to find the optimal policy (Akalin & Loutfi, 2021). This is especially crucial when applied in real-world situations where the number of "live" users is limited by a specific requirement profile for the same (e.g., being an adult with dyslexia) or a special need for protection of the target group, which makes evaluations with real users complex to initiate and implement. Overall, it is worth noting that regardless of the specific type of RL employed (whether model-free or model-based), these methods consistently demonstrated statistical superiority over baselines in half of all applications, which is a promising outcome.

While we observed an increase in the use of DRL methods, an analysis of the reviewed papers regarding their application of DRL (considering only studies that tested for statistical significance) reveals that only 36% of DRL approaches significantly outperformed the baseline. In contrast, 61% of papers utilizing classical RL methods, such as Q-learning or policy iteration, reported a significantly superior RL approach. This suggests that, despite the growing popularity of DRL over the past decade, classical RL approaches still appear to be more effective and sufficient for various educational settings.

RL algorithms can be applied to a wide range of educational domains, from language learning to scientific discovery. The results of our SLR reveal that the variance of the target group and application areas is just as broad as that of the algorithms used. Hence, it can be assumed that different algorithms are suitable for different educational application areas, requiring future research to focus on developing domain-specific algorithms that are tailored to the specific learning goals and challenges of each domain. In this context, another promising direction is to explore the potential of transfer learning in educational RL applications,

involving the development of algorithms that can learn from multiple tasks or domains, and apply that learning to new tasks or domains.

Considered Concepts and Type of Adaptation

Unsurprisingly, when considering the application of RL in an educational context, the vast amount of publications investigated the effect of integrating a RL model on a learning-related outcome variable. Even though only 20% of the reviewed papers were able to demonstrate that at least one RL policy significantly outperformed all baselines, this rises to 51% if only papers with statistical tests for significance are considered, which we consider a promising outcome. Further, a majority of papers evaluating their approach with real learners in regard to the effect on a learning-related outcome variable reported intermediate to large effect sizes, indicating that RL seems to have practical significance for its application to education.

When it comes to the type of adaptation made, we found a higher amount of papers adapting for instructional sequencing (e.g., adjusting the pace of learning based on the student's performance) relative to adaptations for guidance-related aspects (e.g., hint generation or behavior adaptation). It is notable that many papers in the guidance-related category adapted the behavior of their tutoring system to either present a Worked Example or a Problem Solving task (or used a similar approach). We consider this as part of the guidance-related category, because the system does not schedule learning content directly, but instead provides different forms of hints or approaches to handling learning tasks. In general, our classification of the type of adaptation seemed to adequately reflect those made in related work (e.g., Doroudi et al., 2019; Singla et al., 2021).

As suggested by Doroudi et al. (2019), comparing the RL induced policy to relatively weak or random baseline policies increases the chance of yielding better or even significant results. Several algorithms might easily outperform random baselines, although they might not bring additional benefit when compared to expert-crafted learning solutions. However, while overall 35% of reviewed publications applied random baselines, 55% thereof additionally used different baselines for comparison, and of those papers reporting a RL policy significantly outperforming the baseline(s), only 11% compared them to a random baseline only. Thus, it seems that there already is awareness regarding the choice of the baseline policy and how this might affect the results. It is also notable, that random policies might work well when deciding between different beneficial actions, as noted by Doroudi et al. (2019).

Among the papers that utilized expert baselines and reported statistically significant results for an RL approach ($n = 8$), all indicated that their baselines were designed by domain experts, suggesting that in these cases RL was able to capture complex patterns comparable to, and even surpassing, human expertise. RL approaches account for subtle, high-dimensional interactions in the data and might thus be able to discover patterns or decision-making strategies that are not immediately apparent to human experts. If RL performs better, this further implies that

the reward structure effectively aligns with long-term educational success. Considering that the majority of these approaches ($n = 6$) used the students' learning gain or normalized learning gain as a delayed reward, this appears to be an effective reward source.

While half of all papers testing for statistical significance found RL approaches to outperform non-adaptive control conditions, only 14 of the overall 89 studies reviewed included such a baseline in the first place. However, these control conditions are required to effectively assess the effect of adaptive learning environments in isolation and control for confounding variables, such as learning content quality. Thus, non-adaptive control conditions are essential when comparing RL-based systems to establish a clear benchmark and assess the true impact of adaptation. Regarding studies that used RL baselines ($n = 20$), these were often used alongside other baselines, such as expert-designed or random policies. RL baselines frequently consisted of variations of the proposed RL policy (e.g., Huang et al., 2019; Niu & Cao, 2022) or commonly used RL algorithms like DQN and Q-learning.

In summary, half of the reviewed studies that conducted statistical testing reported on RL policies outperforming the baseline. Additionally, while Doroudi et al. (2019) highlighted that many studies rely on weak baselines – raising concerns about the true educational impact of RL-based instructional policies – we observed a growing number of papers applying reasonable baselines, such as expert-designed baselines, or using control conditions without adaptation. This shift is promising, allowing for a more optimistic perspective on the impact of RL in education while suggesting – albeit with caution – that it may have the potential to benefit the learning process of students.

However, regarding the overall objective of applying RL in education – enhancing learning through different approaches – comprehensive insights in how effectively RL methods can actually improve the learning process require comparison to control groups with non-personalized learning environments as well as rigorous methodological approaches to ensure the methodical soundness of evaluations. Over half of all publications included in this review provided no statistical proof for their assumptions, and even though half of the publications included evaluated their systems with learners in a real-world setting, they often lacked longitudinal investigations to determine whether these learning environments are able to benefit learners in the long-term. Thus, the increasing number of publications on applying RL in education is not reflected in an increase in methodological quality of these approaches. This is in line with results of the review from den Hengst et al. (2020) and requires future work to focus not only on developing efficient algorithms, but also on applying reliable measurement techniques and statistical testing while controlling for potential confounding factors to evaluate their effectiveness in actual use contexts. As further noted by Henderson et al. (2018), reproducibility is an issue that should be addressed, because related work on DRL often seems not to be able to reproduce results, reporting varying data for the same baselines (Henderson et al., 2018; Islam et al., 2017).

Best Practices for the Field

Besides overcoming these methodological issues, the area of RL in education holds promise, while at the same time providing potential future research opportunities. When exploring those, several key factors should be considered. In the following section, we thus outline a set of best practices for the research field, summarizing the key insights from our SLR to be applied in practice.

- **Address Research Gaps**: STEM-related subjects appeared to be the most commonly addressed learning topics, whereas other educational areas – such as literacy education and applications for younger learners in general – received comparatively less attention. This requires the development of RL applications for underrepresented target groups (e.g., primary education) and expanding beyond STEM subjects, allowing to scale RL algorithms to large populations of students, while maintaining the effectiveness and personalization of the learning experience.
- **Consider Application Context**: RL methods demonstrated statistically significant superiority over baselines more often for guidance-related tasks, such as providing hints, compared to tasks such as content scheduling. Application context may thus determine the effectiveness of RL methods.
- **Leverage Model-free RL for Adaptive Learning**: Model-free RL methods effectively adapt to educational needs, with half of them showing statistical superiority over baselines. Their ability to learn directly from interactions, combined with ease of implementation and hyperparameter tuning, makes them a strong tool for adaptive learning.
- **Carefully Assess the Need for DRL**: Despite DRL's growing popularity, classical RL methods have shown more consistent effectiveness. The application of established classical RL approaches should be considered first for educational applications, especially for smaller RL problems.
- **Incorporate Learning Gain in RL Reward Functions**: Over half of the papers that reported RL approaches statistically outperforming one or more baselines used learning gain as a source of reward. Thus, students' learning gain, particularly normalized learning gain, should be included in the reward function of RL systems (if applicable), as it effectively quantifies educational progress and aligns well with the goals of enhancing learning outcomes in an educational context.
- **Prioritize Real-World Testing Against Reasonable Baselines**: Real-world RL systems are often challenging to simulate due to the absence of clear-cut rules and conditions, which can lead to artificial constraints when applying them in real-world settings (Dulac-Arnold et al., 2021). Therefore, it is critical to evaluate RL-based educational tools with actual learners before large-scale deployment. While half of studies considered in this SLR do so, this number should increase to ensure practical effectiveness. Further, selecting an appropriate baseline is a critical factor when assessing the effectiveness of RL methods. As Doroudi et al. (2019) note, using weak or random baselines can make RL policies appear more effective or lead to significant results, but not necessarily

meaningful results. Effective control conditions are required to reliably assess the effect of adaptive learning environments Thus, RL-powered agents should be compared against traditional non-adaptive teaching methods and expert-designed systems.

- **Ensure Rigorous Evaluation**: Over half of the reviewed publications lacked statistical validation for their assumptions. As a result, the growing interest in applying RL in education is not matched by improved methodological rigor. It is thus pivotal to conduct statistical tests to reliably assess the effectiveness of RL approaches. Comprehensive statistical testing is essential to provide reliable proof for observed effects.
- **Emphasize Long-Term Learning Outcomes**: While many of the reviewed publications included real-world evaluations, there was a lack of long-term studies. However, short-term interaction does not always lead to students' long-term learning. Longitudinal studies should be carried out more frequently to assess whether RL-driven personalization results in sustained learning benefits.

Limitations of the Review Process

The presented SLR has some limitations. Despite efforts to be comprehensive, the search strategy might have missed relevant studies due to variations in terminology (i.e., combination of keywords used might not have been sufficient), indexing practices, or database coverage. Screening was done by two authors; however, the authors' personal biases and subjective judgments might have influenced the selection, interpretation, and synthesis of studies, which potentially affected the objectivity and reliability of the review results. We aimed to prevent this in advance through applying the PRISMA guidelines (Page et al., 2021), ensuring that our review process is transparent and reproducible.

Further, the focus of the presented SLR on peer-reviewed and published literature potentially excluded relevant findings from grey literature sources and it might thus not be able to comprehensively reflect the actual state of research regarding the application of RL in education. However, considering quality assurance aspects, peer-reviewed publication was deemed as a more important eligibility criterion. In this context, it is also notable that publication bias might have impacted the presented results of this review, i.e., the case where studies with positive or statistically significant results are more likely to be published than those with negative or non-significant findings, and in consequence might have led to an overestimation of the true impact of RL in education. Publication bias is often assessed when conducting a meta-analysis, which we did not perform for the reasons outlined in Sect. 3.4. However, we did review and apply methods commonly used in meta-analyses to detect and quantify publication bias. Several methods have been proposed to investigate publication bias, with the funnel plot combined with Egger's test being the most commonly used (Maier et al., 2022). A funnel plot is a graphical representation that displays effect estimates from individual studies plotted against a measure of

each study's size or precision (Egger et al., 1997; Sterne et al., 2011). Funnel plots are commonly used to evaluate"small-study effects,"referring to whether effect sizes systematically vary between smaller and larger studies (Maier et al., 2022). They are frequently applied to detect publication bias, often in combination with Egger's test.

Given the diverse baselines and outcome measures in this systematic review, along with the limited number of studies providing sufficient data for publication bias analysis, we focused on studies that examined post-test learning performance, comparing an RL approach to an expert-designed baseline, and reported adequate data for analysis, resulting in the inclusion of six studies. We generated a funnel plot and conducted Egger's test, however, given the small number of studies ($n = 6$) that met the criteria for inclusion in this analysis, the interpretation of the funnel plot proved challenging, as no clear pattern was discernible. Moreover, funnel plots are not considered highly reliable for detecting publication bias (Simmonds, 2015). Egger's test also did not indicate significant evidence of publication bias. Recognizing the limitations of these methods, we further applied the precision-effect test and precision-effect estimate with standard errors (PET-PEESE; Stanley & Doucouliagos, 2014), where the PET model also yielded no significant results, indicating no publication bias. Nonetheless, it is important to note that both funnel plot and Egger's test as well as PET-PEESE are sensitive to high heterogeneity among included studies (Maier et al., 2022; Stanley, 2017), which was present in our case, yet we were not able to apply methods accounting for heterogeneity (e.g., selection models) due to the small number of studies ($n < 10$) (Dalton et al., 2016; Maier et al., 2022; Stanley, 2017). Appendix A provides more details on the publication bias analysis.

In summary, while we did not find statistical evidence for publication bias, these results should be interpreted with caution. Given the limited statistical power due to the inclusion of only six studies, the inability to conduct a multi-variate meta-analysis – necessitating a reductionist approach (López-López et al., 2018) – and the high heterogeneity of study results, the presence of publication bias cannot be definitively ruled out. Considering that only six out of 89 studies in this review met the criteria for assessing publication bias, and given that non-significant results are generally less likely to be published or cited (Fanelli, 2010), the possibility of publication bias remains. Thus, while no clear statistical evidence was found, we cannot entirely dismiss its presence in this context.

This further aligns with the relatively small number of papers performing statistical tests. It may suggest that either statistical analysis is not prevalent in this research area or that publication bias might have led to papers primarily reporting results of statistical tests if they found significant differences. This is reflected in the moderate effect sizes found in the majority of papers presenting results of statistical analysis. If there is indeed an approximately average effect of applying RL to education, there should likely be a similar number of both upward and downward spikes of effect sizes. We aimed to address and prevent the given limitations through careful planning and transparent reporting of results, including a more differentiated consideration of reported non-significant results by distinguishing between a lack of statistical significance and the absence of statistical testing.

Conclusion

In our systematic literature review, we provided an overview of the application of Reinforcement Learning in education. RL is a helpful tool to derive optimal strategies to cope with varying environmental conditions. In an educational context, it is able to learn from individual student interactions and can thus provide personalization, regarding both guidance-related (e.g., feedback or hints) or content-related adaptation (e.g., instructional sequencing). Thus, integrating RL methods can benefit learning and the student's affective state.

In accordance with the PRISMA guidelines, we systematically reviewed 89 papers that fit our eligibility criteria and addressed the application of RL in education, reporting on six main outcome domains: The educational context, evaluation strategy, the aim of the study represented by the considered psychological and technical concepts as well as the reported results, the type of algorithm, and the type of adaptation. Overall, there is a notable increase of studies over time, addressing the use of RL in the educational domain, and RL methods have shown promising results in the field of education. This suggests their potential to personalize the learning experience for students and improve the effectiveness of teaching strategies. Over the past six years, there has been a notable rise in the adoption of model-free methods, likely due to their versatile application across different contexts without requiring a model of the environment. However, especially model-free techniques require large amounts of high-quality training data that is often not available for certain educational topics. This requires researchers to rely on simulations, which can be useful for training purposes, but must necessarily be followed by an evaluation with learners in the real world. This further requires rigorous methodological approaches and statistical testing as well as non-adaptive control groups to detect potential effects on relevant dependent variables.

Thus, while the current state of research in the field permits a cautiously optimistic outlook on the impact of RL in education, more research is needed to fully understand the potential of RL in education and to identify the best ways to implement it in practice. Future work should also address current methodological issues, achieved through the implementation of robust study designs and meticulous reporting of results as well as the need for broader and more large-scale deployments and evaluations involving actual users. Based on the insights of our systematic review, we thus derived best practices for the field to address and overcome these shortcomings, summarizing the key insights from our SLR to be applied in practice while addressing methodological, technical, and procedural issues. Despite these challenges, the use of RL in education is an active and growing area of research, and we can expect to see continued progress and improvement in this field in the coming years.

Appendix A: Details on the Analysis of Publication Bias

This section provides more details on the conducted publication bias analysis. Given the diverse baselines and outcome measures in this systematic review, along with the limited number of studies providing sufficient data for publication bias analysis, we focused on studies that examined post-test learning performance, comparing an RL approach to an expert-designed baseline, and reported adequate data for analysis. This resulted in the inclusion of six studies: Ausin et al. (2021), Ausin et al. (2022), Chi et al. (2014), Hostetter et al. (2023c), Jung et al. (2024), and Riedmann et al. (2024). All publication bias analysis was conducted with JASP (JASP Team, 2021). Interpreting the funnel plot proved challenging, lacking a recognizable pattern due to the small number of studies, and in general, funnel plots are not considered highly reliable for detecting publication bias (Simmonds, 2015). Egger's test also showed no significant indication of publication bias ($z = -0.20$, $p = .842$). However, both methods rely on the assumption that smaller studies with large effect estimates are more likely to be published, with larger studies being less influenced by this bias, while, in practice, publication bias may not always follow this pattern (Maier et al., 2022; Sterne et al., 2001). To account for this, we additionally applied the precision-effect test and precision-effect estimate with standard errors (PET-PEESE; Stanley & Doucouliagos, 2014), where the PET model also yielded no significant results ($t(4) = 1.34$, $p = .251$), indicating no publication bias.

Nonetheless, it is important to note that both funnel plot and Egger's test as well as PET-PEESE are sensitive to high heterogeneity among included studies (Maier et al., 2022; Stanley, 2017), which was present in our case, as indicated by tests of heterogeneity ($Q(5) = 1407.74$, $p < .001$, $I^2 > 75\%$). Further, these methods as well as other approaches accounting for heterogeneity (e.g., selection models) perform poorly when applied to meta-analyses with a small number of studies ($n < 10$) (Dalton et al., 2016; Maier et al., 2022; Stanley, 2017), making it difficult to reliably assess publication bias due to limited statistical power. In summary, while we did not find statistical evidence for publication bias, these results should be interpreted with caution. Given the limited statistical power due to the inclusion of only six studies, the inability to conduct a multi-variate meta-analysis – necessitating a reductionist approach (López-López et al., 2018) – and the high heterogeneity of study results, the presence of publication bias cannot be definitively ruled out. Considering that only six out of 89 studies in this review met the criteria for assessing publication bias, and given that non-significant results are generally less likely to be published or cited (Fanelli, 2010), the possibility of publication bias remains. Thus, while no clear statistical evidence was found, we cannot entirely dismiss its presence in this context.

Appendix B: Overview of All Included Articles Organized According to the Specified Data Items for this Systematic Literature Review

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
1	Abdelshiheed et al., 2023a	DoubleDQN	NoAdaptation	Math	Higher education	Live	G	LP* (normalized learning gain)
2	Ai et al., 2019	TRPO	Heuristic	Math	Elementary school	Real interaction data	C	LP† (student's knowledge level), LE† (accelerate learning process)
3	Alam et al., 2024	DoubleDQN	Random, RL, NoAdaptation	Math	Higher education	Live	G	LP (post-test performance), LE (normalized learning gain per total tutor time), TT
4	Ausin et al., 2019	DQN, DoubleDQN	Random	Math	Higher education	Live	G	LE (post-test efficiency)
5	Ausin et al., 2021	DQN, DuelingDQN	Expert, NoAdaptation	Math	Higher education	Live	G	LP (post-test performance), LE (post-test score by training time)
6	Ausin et al., 2022	DQN, DoubleDQN	Expert, NoAdaptation	Math	Higher education	Live	G	LP (post-test score), TT
7	Barnes et al., 2008	Value iteration	NoAdaptation	Other	Higher education	Live	G	TC† (percentage of students completing the given tasks)
8	Bassen et al., 2020	PPO	Expert, NoAdaptation	Math	Higher education	Live	C	LP* (learning gain), TC* (dropout rate)
9	Caro et al., 2023	Q-learning	NoAdaptation	Other	High school	Simulated data	C	LP† (cognitive growth)
10	Chakraborty et al., 2021	Other	NoAdaptation	Math	High school	Live	C	LP (test scores)
11	Chen et al., 2020	Q-learning	Other	Language	Elementary school	Live	G	LP* (vocabulary acquisition), EM* (affective engagement)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
12	Chi et al., 2010	Policy iteration	Random, RL	Physics	Higher education	Live	G	LP* (post-test and normalized learning gain)
13	Chi et al., 2014	Policy iteration	Expert	Physics	Higher education	Live	G	LP* (post-test and normalized learning gain), TT
14	Efremov et al., 2020	REINFORCE	Random, Other	Computer-related	Not specified	Real interaction data	G	LE† (hint accuracy)
15	Fernandes et al., 2023	Q-learning	Not specified	Not specified	Not specified	Simulated data	C	AP† (F1, recall, and precision score)
16	Flores et al., 2019	Q-learning	Not specified	Other	Not specified	Live	C	LP† (learning session score), AP† (precision and recall score)
17	Fotopoulou et al., 2020	SLATEQ	Not specified	Other	Not specified	Simulated data	C	AP† (average episode reward and Huber loss)
18	Gao et al., 2023a	DQN	Expert	Math	Higher education	Real interaction data	G	AP* (absolute error, regret, and rank correlation coefficient)
19	Georgila et al., 2019	LSPI	Heuristic	Other	Higher education	Live	G	LP (learning gain)
20	Gkatzia et al., 2013	Tabular TD learning	Random, Expert, RL	Computer-related	Higher education	Real interaction data	G	AP* (RL reward)
21	Gong et al., 2023	REINFORCE	Other	Not specified	Adults	Real interaction data	C	AP† (Mean Reciprocal Rank, Normalized Discounted Cumulative Gain, and Hit Ratio of top-K items)
22	Gordon et al., 2016	SARSA	Expert	Language	Preschool	Live	G	LP (Spanish words learned), EM (valence, engagement)
23	Han et al., 2020	Actor-critic	Random	Not specified	Not specified	Simulated data	C	LP† (score)
24	Hostetter et al., 2023a	DQN	Expert, NoAdaptation	Math	Higher education	Live	G	LP* (post-test and normalized learning gain)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
25	Hostetter et al., 2023c	Fuzzy Conservative Q-learning	Expert	Math	Higher education	Live	G	LP* (post-test score)
26	Huang et al., 2019	DQN-based	Random, RL	Math	High school	Real interaction data	C	AP† (Mean Average Precision, Normalized Discounted Cumulative Gain of top-K items, F1 score)
27	Iglesias et al., 2009a	Q-learning	Not specified	Not specified	Not specified	Simulated data	C	AP† (algorithm learning curves)
28	Iglesias et al., 2009b	Q-learning	Expert	Not specified	Higher education	Live	C	LP† (student's level of knowledge), LE† (course pages required to learn the course content), TT†
29	Intayoad et al., 2018	SARSA	Not specified	Not specified	Not specified	Real interaction data	C	AP† (Root Mean Square Error)
30	Islam et al., 2021	Q-learning	RL	Not specified	Not specified	Real interaction data	C	LP† (student's learning rate)
31	Jeewantha et al., 2021	Q-learning	Random	Language	Adults	Not specified	C	LP† (accuracy of speaking and writing activities)
32	Ju et al., 2021	DQN	RL	Math	Higher education	Live	G	LP* (post-test and normalized learning gain), TT*
33	Jung et al., 2024	DoubleDQN, Dreamer	Expert	Math	Higher education	Live	G	LP (post-test and normalized learning gain), TT (training time in minutes)
34	Kakdas et al., 2024	Q-learning	RL	Other	Higher education	Live	C	LE† (efficiency of task completion)
35	Kim et al., 2021	TRPO	Random	Computer-related	Adults	Real interaction data	C	AP† (five-fold cross validation)
36	Koroveshi & Ktona, 2021	DQN	Not specified	Computer-related	Not specified	Simulated data	C	AP† (RL reward received per episode)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
37	Li et al., 2023b	DQN	Random	Not specified	Not specified	Simulated data	C	AP† (Root Mean Square Error)
38	Li et al., 2024b	PPO	RL, Other	Math	Not specified	Real interaction data	C	LP† (learners' mastery of target knowledge points)
39	Liang & Yu, 2024	Other	Other	Not specified	Not specified	Real interaction data	C	AP† (Mean Reciprocal Rank, Hits@1, Root Mean Square Error)
40	Liu et al., 2018	DQN	RL, NoAdaptation	Not specified	Not specified	Live	C	LP† (learning performance)
41	Liu et al., 2019	Actor-critic	Random, RL, Other	Not specified	Not specified	Real interaction data	C	LP† (learners' knowledge level)
42	Malpani et al., 2011	Actor-critic	Random	Not specified	Not specified	Simulated data	C	LP† (maximize students' learning)
43	Mandel et al., 2014	QMDP	Random, Expert	Math	Not specified	Live	C	AP* (average RL reward)
44	Martin & Arroyo, 2004	Policy iteration	Not specified	Not specified	Not specified	Simulated data	G	LP† (student performance), LE† (efficiency of hint sequencing)
45	Ming & Hua, 2010	Other	RL	Not specified	Higher education	Real interaction data	C	AP† (algorithm efficiency, scheduling accuracy, and largest debugging rate)
46	Mohana et al., 2024	Other	Other	Language	Not specified	Real interaction data	G	AP† (precision, recall, and F1 score)
47	Niu & Cao, 2022	Q-learning	RL	Not specified	Not specified	Simulated data	C	AP† (effectiveness and optimality of reward scheme)
48	Oralbayeva et al., 2022	Q-learning	Random	Language	Higher education	Live	C	LP (learned letters)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
49	Orsoni et al., 2023	PPO, A2C	Random	Math	Higher education	Real interaction data	C	AP† (Akaike Information Criterion, Bayesian Information Criterion, and Log Likelihood)
50	Oyuga Anne & Maina, 2021	Q-learning	Not specified	Not specified	Not specified	Simulated data	C	AP† (RL reward)
51	Pan & Yang, 2024	Q-learning	RL	Other	Higher education	Simulated data	C	AP† (accuracy, efficiency, and safety)
52	Park et al., 2019	Q-learning	NoAdaptation	Other	Elementary school	Live	C	LP* (vocabulary test), EM* (engagement)
53	Pérez et al., 2024	Algorithm based on Q-learning	RL, Other	Math	Middle school	Real interaction data	C	LP† (learning gain), AP† (utility, skill selection rate)
54	Pietquin et al., 2011	LSPI	Random, Other	Language	Not specified	Simulated data	C	AP† (cumulative RL reward)
55	Pögelt et al., 2024	PPO, DQN	Random	Math	Higher education	Simulated data	C	AP† (average RL reward)
56	Pu et al., 2020	DQN	Random	Math	Middle school	Live	C	LE† (mean score difference to length of sequences)
57	Raghuveer et al., 2014	CRBL	Not specified	Computer-related	Not specified	Live	C	LP† (learning outcomes for application skills)
58	Ravari et al., 2021	Q-learning	NoAdaptation	Geology	Adults	Live	G	EM* (engagement)
59	Reddy et al., 2017	TRPO	Random, Heuristic	Not specified	Not specified	Simulated data	C	AP† (expected recall likelihood and log-likelihood)
60	Riedmann et al., 2024	DQN	Expert	Language	Elementary school	Live	G	LP (reading performance), EM (motivation)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
61	Rowe & Lester, 2015	Value iteration	Random	Biology	Middle school	Live	G	LP (microbiology content knowledge)
62	Ruan et al., 2024	PPO	NoAdaptation	Math	Elementary school	Live	G	LP (pre-post score improvement), EM (engagement)
63	Sawyer et al., 2017	Convex Hull Value iteration	Other	Biology	Middle school	Real interaction data	G	LP* (normalized learning gain), EM* (Presence Questionnaire)
64	Shawky & Badawi, 2018	Other	Not specified	Not specified	Not specified	Simulated data	C	LP† (learning outcomes), EM† (satisfaction)
65	Shen et al., 2018a	LSPI	Random, RL	Math	Higher education	Live	G	LP (post-test score and normalized learning gain), TT
66	Shen et al., 2018b	Other	Random	Math	Higher education	Live	G	LP (transfer post-test score)
67	Shen et al., 2018c	Q-learning	Random, RL	Math	Higher education	Live	G	LP (post-test score)
68	Shen et al., 2021	Q-learning	Not specified	Math	Not specified	Simulated data	C	AP† (give appropriate level problems)
69	Shin & Bulut, 2022	Actor-critic	Not specified	Language	Elementary school	Real interaction data	C	AP† (accuracy, F1 score)
70	Spain et al., 2021	Value iteration	Not specified	Other	Adults	Real interaction data	G	LP† (learning gains)
71	Stamper et al., 2008	Value iteration	Not specified	Other	Higher education	Live	G	AP† (ability to give hints when requested)
72	Su et al., 2013	Q-learning	Heuristic	Language	Adults	Real interaction data	C	LP† (average scores of pronunciation units)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
73	Sun et al., 2024	Q-learning	Expert	Other	Not specified	Simulated data	C	LP† (higher accuracy in sports training movements), EM† (user satisfaction), TT†
74	Tang et al., 2019	Q-learning	Random	Not specified	Not specified	Simulated data	C	LP† (total number of skills mastered)
75	Tetrault & Litman, 2006a	Policy iteration	Not specified	Physics	Higher education	Real interaction data	G	LP† (highest learning gain)
76	Vassoyan et al., 2023	REINFORCE	Random, RL	Not specified	Not specified	Simulated data	C	AP† (episodic return)
77	Wan et al., 2023	DoubleDQN	NoAdaptation, Other	Not specified	Not specified	Real interaction data	C	LP† (students' average knowledge state)
78	Wang, 2014b	Other	Random	Computer-related	Not specified	Live	G	AP* (rejection rate)
79	Wang et al., 2023	DQN, DoubleDQN, DuelingDQN	Not specified	Not specified	Not specified	Real interaction data	C	AP† (Q-value, return value)
80	Wang & Song, 2024	Q-learning	Not specified	Not specified	Not specified	Simulated data	C	LP† (quiz accuracy), LE† (improvement in the average time and accuracy), TT†, AP† (system response time)
81	Whitehill & Movellan, 2018	Other	Expert	Language	Adults	Live	C	LE (average time to pass the test), TC (task completion)
82	Wu et al., 2024	DoubleDQN	Other	Not specified	Not specified	Live	C	LP† (student learning ability)

	Author, year	Type of algorithm	Baseline(s)	Subject of the learning environment	Educational target group	Evaluation strategy	Type of adaptation	Aim of study, with * for $p < .05$, † for no statistical comparison, details in parentheses
83	Yantao & Wei, 2024	Q-learning	Not specified	Not specified	Not specified	Simulated data	G	TC† (student participation)
84	Yessad, 2023	Q-learning	Random, Expert	Not specified	Not specified	Simulated data	C	LE* (number of activities proposed to students before acquiring all knowledge components)
85	Zhang & Goh, 2019	BPG	Random, RL, Other	Not specified	Not specified	Simulated data	C	LP† (probability of a user answering a task correctly)
86	Zhang & Goh, 2021	BPG	Random, Expert	Other	Adults	Live	C	TT (memorization time)
87	Zhang, 2023	Q-learning	Expert	Language	Middle school	Live	G	LP† (English learning level), EM† (learning motivation)
88	Zhou et al., 2019	Policy iteration	Random, RL	Math	Higher education	Live	G	LP* (post-test and normalized learning gain), TT*
89	Zhou et al., 2020	Other	Random	Math	Higher education	Live	G	LP* (post-test and normalized learning gain), EM* (engagement through student-system interaction logs), TT*

Entries labelled as 'G' describe the guidance-related category, while 'C' comprises all publications with content-related adaptations. Regarding the aim of the study, 'LP' denotes learning/performance, 'EM' for engagement/motivation, 'LE' for learning efficiency, 'TT' for time on task, 'AP' for algorithm performance, and 'TC' for task completion.

Acknowledgments We thank the reviewers and especially Shayan Doroudi for the detailed and comprehensive feedback on our manuscript.

Author Contributions A.R. was the main author responsible for conducting literature research, methodology definition, and paper writing, supervised by B.L. P.S. was the second reviewer in the screening phase and has been involved in structuring and writing the paper. A.R. wrote the first draft of the manuscript, and P.S. and B.L. commented on scope, structure, and content of previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

Registration Information This review was not registered. The protocol for the systematic review is available upon request from the corresponding author.

References

*Indicates a study included in the systematic literature review

Abdelshiheed, M., Hostetter, J. W., Barnes, T., & Chi, M. (2023b). Bridging Declarative, Procedural, and Conditional Metacognitive Knowledge Gap Using Deep Reinforcement Learning. In *CogSci'23: The 45th Annual Conference of the Cognitive Science Society.*

*Abdelshiheed, M., Hostetter, J. W., Barnes, T., spsampsps Chi, M. (2023a). Leveraging deep reinforcement learning for metacognitive interventions across intelligent tutoring systems. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, spsampsps V. Dimitrova (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 13916. Artificial Intelligence in Education: 24th International Conference, AIED 2023, Tokyo, Japan, July 3–7, 2023, Proceedings* (Vol. 13916, pp. 291–303). Springer Nature Switzerland; Imprint Springer. https://doi.org/10.1007/978-3-031-36272-9_24

Afsar, M. M., Crump, T., & Far, B. (2023). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys, 55*(7), 1–38. https://doi.org/10.1145/3543846

*Ai, F., Chen, Y., Guo, Y., Zhao, Y., Wang, Z., Fu, G., & Wang, G. Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 240–245).

Akalin, N., & Loutfi, A. (2021). Reinforcement learning approaches in social robotics. *Sensors*, *21*(4). https://doi.org/10.3390/s21041292

Akanksha, E., Jyoti, Sharma, N., & Gulati, K. (2021). Review on reinforcement learning, research evolution and scope of application. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1416–1423). IEEE. https://doi.org/10.1109/ICCMC51019.2021.9418283

*Alam, N., Mostafavi, B., Tithi, S. D., Chi, M., & Barnes, T. (2024). How Much Training is Needed? Reducing Training Time using Deep Reinforcement Learning in an Intelligent Tutor. In *Proceedings of the 17th International Conference on Educational Data Mining*.

Amin, S., Uddin, M. I., Alarood, A. A., Mashwani, W. K., Alzahrani, A., & Alzahrani, A. O. (2023). Smart e-learning framework for personalized adaptive learning and sequential path recommendations using reinforcement learning. *IEEE Access, 11*, 89769–89790. https://doi.org/10.1109/ACCESS.2023.3305584

*Ausin, M. S., Azizsoltani, H., Barnes, T., & Chi, M. (2019). Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)* (pp. 168–177).

*Ausin, M. S., Maniktala, M., Barnes, T., spsampsps Chi, M. (2021). Tackling the credit assignment problem in reinforcement learning-induced pedagogical policies with neural networks. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, spsampsps V. Dimitrova (Eds.), *Lecture Notes in Computer Science. ARTIFICIAL INTELLIGENCE IN EDUCATION: 22nd international conference, aied 2021* (Vol. 12748, pp. 356–368). SPRINGER NATURE. https://doi.org/10.1007/978-3-030-78292-4_29

Ausin, M. S., Maniktala, M., Barnes, T., & Chi, M. (2022). The impact of batch deep reinforcement learning on student performance: A simple act of explanation can go a long way. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-022-00312-3

Balaji, B., Mallya, S., Genc, S., Gupta, S., Dirac, L., Khare, V., Roy, G., Sun, T., Tao, Y., Townsend, B., Calleja, E., Muralidhara, S., & Karuppasamy, D. (2020). Deepracer: Autonomous racing platform for experimentation with sim2real reinforcement learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 2746–2754). IEEE. https://doi.org/10.1109/ICRA40945.2020.9197465

*Barnes, T., Stamper, J., Lehman, L., & Croy, M. (2008). A pilot study on logic proof tutoring using hints generated from historical student data. In *Proceedings of Educational Data Mining 2008, The 1st International Conference on Educational Data Mining, Montreal, Québec, Canada, June 20–21, 2008*.

Barnes, T., spsampsps Stamper, J. (2008). Toward automatic hint generation for logic proof tutoring using historical student data. In B. P. Woolf, B. Woolf, E. Aïmeur, R. Nkambou, spsampsps S. Lajoie (Eds.), *Lecture Notes in Computer Science: Vol. 5091. Intelligent tutoring systems: 9th international conference, ITS 2008, Montreal, Canada, June 23 - 27, 2008; proceedings* (Vol. 5091, pp. 373–382). Springer. https://doi.org/10.1007/978-3-540-69132-7_41

*Bassen, J., Balaji, B., Schaarschmidt, M., Thille, C., Painter, J., Zimmaro, D., Games, A., Fast, E., & Mitchell, J. C. (2020). Reinforcement learning for the adaptive scheduling of educational activities. In R. Bernhaupt (Ed.), *ACM Digital Library, Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376518

Beck, J. (1997). Modeling the student with reinforcement learning. In *Machine learning for User Modeling Workshop at the Sixth International Conference on User Modeling*.

Beck, J. E. (1998). Learning to teach with a reinforcement learning agent. In J. Mostow & C. Rich (Eds.), *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98, IAAI 98, July 26–30, 1998, Madison, Wisconsin, USA* (p. 1185). AAAI Press / The MIT Press. http://www.aaai.org/Library/AAAI/1998/aaai98-181.php

Beck, J. E., Woolf, B. P., & Beal, C. R. (2000). ADVISOR: A machine learning architecture for intelligent tutor construction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence* (pp. 552–557).

Beck, J. E., spsampsps Woolf, B. P. (2000). High-level student modeling with machine learning. In G. Gauthier (Ed.), *Lecture Notes in Computer Science: Vol. 1839. Intelligent tutoring systems: 5th international conference; proceedings* (Vol. 1839, pp. 584–593). Springer. https://doi.org/10.1007/3-540-45108-0_62

Belfer, R., Kochmar, E., spsampsps Serban, I. V. (2022). Raising student completion rates with adaptive curriculum and contextual bandits. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, spsampsps V. Dimitrova (Eds.), *Lecture Notes in Computer Science: Vol. 13355. Artificial Intelligence in Education: 23rd International Conference, AIED 2022, Durham, UK, July 27–31, 2022, Proceedings, Part I* (1st ed. 2022, Vol. 13355, pp. 724–730). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-031-11644-5_74

Bellman, R. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences of the United States of America, 38*(8), 716–719. https://doi.org/10.1073/pnas.38.8.716

Bellotti, F., Berta, R., de Gloria, A., & Primavera, L. (2009). Adaptive experience engine for serious games. *IEEE Transactions on Computational Intelligence and AI in Games, 1*(4), 264–280. https://doi.org/10.1109/TCIAIG.2009.2035923

Bennane, A. (2013). Adaptive educational software by applying reinforcement learning. *Informatics in Education*, *12*(1), 13–28. https://doi.org/10.15388/infedu.2013.02

Bittencourt, I., Tadeu, M., & Costa, E. (2006). Combining AI techniques into a legal agent-based intelligent tutoring system. In *Eighteenth International Conference on Software Engineering and Knowledge Engineering-SEKE* (Vol. 18, pp. 35–40).

Boussakssou, M., Hssina, B., & Erittali, M. (2020). Towards an adaptive e-learning system based on q-learning algorithm. *Procedia Computer Science, 170*, 1198–1203. https://doi.org/10.1016/j.procs.2020.03.028

*Caro, M. F., Quitian, L., Giraldo, J. C., & Lengua-Cantero, C. (2023). A Formal Model for Personalized Learning Path using Artificial Intelligence for Instructional Planning with a Focus on 21st-Century Skills and Environmental Awareness. In *2023 IEEE Colombian Caribbean Conference (C3)* (pp. 1–6). IEEE. https://doi.org/10.1109/C358072.2023.10436195

*Chakraborty, N., Roy, S., Leite, W. L., Faradonbeh, M. K. S., & Michailidis, G. (2021). The effects of a personalized recommendation system on students' high-stakes achievement scores: A field experiment. In *Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021)* (pp. 588–594).

Chen, H., Park, H. W., & Breazeal, C. (2020). Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement. *Computers & Education, 150*, Article 103836. https://doi.org/10.1016/j.compedu.2020.103836

Chi, M., Jordan, P., VanLehn, K., & Hall, M. (2008). Reinforcement learning based feature selection for developing pedagogically effective tutorial dialogue tactics. In *Proceedings of Educational Data Mining 2008 - 1st International Conference on Educational Data Mining* (pp. 258–265).

*Chi, M., VanLehn, K., Litman, D., spsampsps Jordan, P. (2010). Inducing effective pedagogical strategies using learning context features. In P. de Bra, P. Del Brassey, A. Kobsa, spsampsps D. Chin (Eds.), *Lecture Notes in Computer Science / Information Systems and Applications, incl. Internet/Web, and HCI: Vol. 6075. User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, June 20–24, 2010 ; proceedings* (Vol. 6075, pp. 147–158). Springer. https://doi.org/10.1007/978-3-642-13470-8_15

*Chi, M., Jordan, P., spsampsps VanLehn, K. (2014). When is tutorial dialogue more effective than step-based tutoring? In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, A. Kobsa, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, D. Terzopoulos, D. Tygar, G. Weikum, S. Trausan-Matu, K. E. Boyer, M. Crosby, K. Panourgia, spsampsps Ş. Trăuşan-Matu (Eds.), *Lecture Notes in Computer Science: Vol. 8474. Intelligent tutoring systems: 12th international conference, ITS 2014, Honolulu, HI, USA, June 5 - 9, 2014; proceedings* (Vol. 8474, pp. 210–219). Springer. https://doi.org/10.1007/978-3-319-07221-0_25

Chi, M., VanLehn, K., Litman, D., & Jordan, P. (2011). Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Modeling and User-Adapted Interaction, 21*(1–2), 137–180. https://doi.org/10.1007/s11257-010-9093-1

Clement, B., Roy, D., Oudeyer, P.-Y., & Lopes, M. (2015). Multi-armed bandits for intelligent tutoring systems. *Journal of Educational Data Mining, 7*(2), 20–48. https://doi.org/10.5281/ZENODO.3554667

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Erlbaum. https://doi.org/10.4324/9780203771587

Condor, A., & Pardos, Z. (2022). A deep reinforcement learning approach to automatic formative feedback. In *Proceedings of the 15th International Conference on Educational Data Mining* (pp. 662–666). https://doi.org/10.5281/zenodo.6853061

Croy, M., Barnes, T., & Stamper, J. (2008). Towards an intelligent tutoring system for propositional proof construction. In *Proceedings of the 2008 conference on Current Issues in Computing and Philosophy*.

Cui, L. (2023). Research of Intelligent Tutoring System Based on Deep Learning Computer Technology. In *2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICAISC58445.2023.10199414

Dalton, J. E., Bolen, S. D., & Mascha, E. J. (2016). Publication Bias: The Elephant in the Review. *Anesthesia and Analgesia, 123*(4), 812–813. https://doi.org/10.1213/ANE.0000000000001596

den Hengst, F., Grua, E. M., el Hassouni, A., & Hoogendoorn, M. (2020). Reinforcement learning for personalization: A systematic literature review. *Data Science, 3*(2), 107–147. https://doi.org/10.3233/DS-200028

Dorça, F. A., Lima, L. V., Fernandes, M. A., & Lopes, C. R. (2013). Comparing strategies for modeling students learning styles through reinforcement learning in adaptive and intelligent educational systems: An experimental analysis. *Expert Systems with Applications, 40*(6), 2092–2101. https://doi.org/10.1016/j.eswa.2012.10.014

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward? *International Journal of Artificial Intelligence in Education, 29*(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning, 110*(9), 2419–2468. https://doi.org/10.1007/s10994-021-05961-4

*Efremov, A., Ghosh, A., & Singla, A. K. (2020). Zero-shot learning of hint policy via reinforcement learning and program synthesis. In *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)* (pp. 388–394).

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ : British Medical Journal, 315*(7109), 629–634. https://doi.org/10.1136/bmj.315.7109.629

El Fazazi, H., Elgarej, M., Qbadou, M., & Mansouri, K. (2021). Design of an adaptive e-learning system based on multi-agent approach and reinforcement learning. *Engineering, Technology & Applied Science Research, 11*(1), 6637–6644. https://doi.org/10.48084/etasr.3905

Fahad Mon, B., Wasfi, A., Hayajneh, M., Slim, A., & Abu Ali, N. (2023). Reinforcement learning in education: A literature review. *Informatics, 10*(3), 74. https://doi.org/10.3390/informatics10030074

Fahid, F. M., Rowe, J., Kim, Y., Srivastava, S., & Lester, J. (2024). Online reinforcement learning-based pedagogical planning for narrative-centered learning environments. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(21), 23191–23199. https://doi.org/10.1609/aaai.v38i21.30365

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS ONE, 5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Fenza, G., Orciuoli, F., & Sampson, D. G. (2017). Building adaptive tutoring model using artificial neural networks and reinforcement learning. In M. Chang (Ed.), *Icalt 2017: Ieee 17th International Conference on Advanced Learning Technologies: Proceedings: 3–7 July 2017, Timişoara, Romania.* IEEE. https://doi.org/10.1109/ICALT.2017.124

Ferguson, K., Woolf, B. P., & Mahadevan, S. (2009). Transfer learning and representation discovery in intelligent tutoring systems. *Frontiers in Artificial Intelligence and Applications, 200*, 605–607. https://doi.org/10.3233/978-1-60750-028-5-605

*Fernandes, C. W., Miari, T., Rafatirad, S., & Sayadi, H. (2023). Unleashing the Potential of Reinforcement Learning for Enhanced Personalized Education. In *2023 IEEE Frontiers in Education Conference (FIE)* (pp. 1–5). IEEE. https://doi.org/10.1109/FIE58773.2023.10342902

*Flores, A., Alfaro, L., & Herrera, J. (2019). Proposal model for e-learning based on case based reasoning and reinforcement learning. In C. d. Rocha Brito & M. M. Ciampi (Eds.), *Modern educational paradigms for computer and engineering career: Proceedings: Edunine2019 - III IEEE World*

Engineering Education Conference : March 17 to 19, 2019, Lima, Peru (pp. 1–6). IEEE. https://doi.org/10.1109/EDUNINE.2019.8875800

*Fotopoulou, E., Zafeiropoulos, A., Feidakis, M., Metafas, D., spsampsps Papavassiliou, S. (2020). An interactive recommender system based on reinforcement learning for improving emotional competences in educational groups. In V. Kumar spsampsps C. Troussas (Eds.), *Programming and Software Engineering: Vol. 12149. Intelligent Tutoring Systems: 16th International Conference, ITS 2020, Athens, Greece, June 8–12, 2020, Proceedings* (1st ed. 2020, Vol. 12149, pp. 248–258). Springer International Publishing; Imprint: Springer. https://doi.org/10.1007/978-3-030-49663-0_29

El Fouki, M., Aknin, N., & El. Kadiri, K. E. (2017). Intelligent adapted e-learning system based on deep reinforcement learning. In J. Zbitou (Ed.), *ACM Digital Library, Proceedings of the 2nd International Conference on Computing and Wireless Communication Systems* (pp. 1–6). ACM. https://doi.org/10.1145/3167486.3167574

François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., & Pineau, J. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, *11*(3–4), 219–354. https://doi.org/10.1561/2200000071

Frej, J., Shah, N., Knezevic, M., Nazaretsky, T., & Käser, T. (2024). Finding Paths for Explainable MOOC Recommendation: A Learner Perspective. In *Proceedings of the 14th Learning Analytics and Knowledge Conference* (pp. 426–437). ACM. https://doi.org/10.1145/3636555.3636898

Frenoy, R., Soullard, Y., Thouvenin, I., & Gapenne, O. (2016). Adaptive training environment without prior knowledge. In J. Vassileva (Ed.), *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 131–139). ACM. https://doi.org/10.1145/2930238.2930256

Gao, Y., Barendregt, W., & Castellano, G. (2017). Personalised human-robot co-adaptation in instructional settings using reinforcement learning. In *IVA Workshop on Persuasive Embodied Agents for Behavior Change: PEACH 2017, August 27, Stockholm, Sweden*.

Gao, Y., Barendregt, W., Obaid, M., & Castellano, G. (2018). When robot personalisation does not help: Insights from a robot-supported learning study. In J.-J. Cabibihan (Ed.), *Ieee RO-MAN 2018: The 27th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 705–712). IEEE. https://doi.org/10.1109/ROMAN.2018.8525832

*Gao, G., Ju, S., Ausin, M. S., & Chi, M. (2023a). HOPE: Human-centric off-policy evaluation for e-learning and healthcare. In *AAMAS '23, Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1504–1513).

Gao, H., Zeng, Y., Ma, B., & Pan, Y. (2023b). *Improving knowledge learning through modelling students' practice-based cognitive processes*. Advance online publication. https://doi.org/10.1007/s12559-023-10201-z

*Georgila, K., Core, M. G., Nye, B. D., Karumbaiah, S., Auerbach, D., & Ram, M. (2019). Using reinforcement learning to optimize the policies of an intelligent tutoring system for interpersonal skills training. In *AAMAS '19, Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 737–745).

Geramifard, A., Dann, C., Klein, R. H., Dabney, W., & How, J. P. (2015). RLPy: A value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research, 16*(1), 1573–1578.

*Gkatzia, D., Hastie, H., Janarthanam, S., & Lemon, O. (2013). Generating student feedback from time-series data using reinforcement learning. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 115–124).

Gong, J., Wan, Y., Liu, Y., Li, X., Zhao, Y., Wang, C., Lin, Y., Fang, X., Feng, W., Zhang, J., & Tang, J. (2023). Reinforced MOOCs concept recommendation in heterogeneous information networks. *ACM Transactions on the Web, 17*(3), 1–27. https://doi.org/10.1145/3580510

*Gordon, G., Spaulding, S., Westlund, J. K., Lee, J. J., Plummer, L., Martinez, M., Das, M., & Breazeal, C. (2016). Affective personalization of a social robot tutor for children's second language skills. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), 3951–3957. https://ojs.aaai.org/index.php/aaai/article/view/9914

Graesser, A. C., Conley, M. W., spsampsps Olney, A. (2012). Intelligent tutoring systems. In K. R. Harris, S. Graham, spsampsps T. C. Urdan (Eds.), *APA handbooks in psychology. APA educational psychology handbook* (1st ed., pp. 451–473). American Psychological Association. https://doi.org/10.1037/13275-018

Guran, A.-M., Cojocar, G.-S., spsampsps Dioşan, L.-S. (2021). Towards smart edutainment applications for young children: A proposal. In A. I. Cristea spsampsps C. Troussas (Eds.), *Springer eBook*

Collection: Vol. 12677. Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings (1st ed. 2021, Vol. 12677, pp. 439–443). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-80421-3_48

Haider, S. A., Ahmad, K. M., Zahid, A., AlGhamdi, A., Keshta, I., & Soni, M. (2024). Genetic and Deep Reinforcement Learning-Based Intelligent Course Scheduling for Smart Education. In *Proceedings of the 2024 International Conference on Artificial Intelligence and Teacher Education* (pp. 117–124). ACM. https://doi.org/10.1145/3702386.3702398

Han, R., Chen, K., & Tan, C. (2020). Curiosity-driven recommendation strategy for adaptive learning via deep reinforcement learning. *The British Journal of Mathematical and Statistical Psychology, 73*(3), 522–540. https://doi.org/10.1111/bmsp.12199

Hare, R., & Tang, Y. (2023). Reinforcement Learning with Experience Sharing for Intelligent Educational Systems. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 1431–1436). IEEE. https://doi.org/10.1109/SMC53992.2023.10394095

Hare, R., Tang, Y., & Ferguson, S. (2024). An Intelligent Serious Game for Digital Logic Education to Enhance Student Learning. *IEEE Transactions on Education, 67*(3), 387–394. https://doi.org/10.1109/TE.2024.3359001

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In *AAAI Conference On Artificial Intelligence (AAAI)* (pp. 3207–3214).

*Hostetter, J. W., Abdelshiheed, M., Barnes, T., & Chi, M. (2023c). A self-organizing neuro-fuzzy q-network: Systematic design with offline hybrid learning. In *AAMAS '23, Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (pp. 1248–1257).

Hostetter, J., Conati, C., Yang, X., Abdelshiheed, M., Barnes, T., & Chi, M. (2023b). XAI to increase the effectiveness of an intelligent pedagogical agent. In *ACM International Conference on Intelligent Virtual Agents (IVA2023)*.

*Hostetter, J., Abdelshiheed, M., Barnes, T., & Chi, M. (2023a). Leveraging fuzzy logic towards more explainable reinforcement learning-induced pedagogical policies on intelligent tutoring systems. In *IEEE International Conference on Fuzzy Systems (FUZZ 2023)*.

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Press.

*Huang, Z., Liu, Q., Zhai, C., Yin, Y., Chen, E., Gao, W., & Hu, G. (2019). Exploring multi-objective exercise recommendations in online education systems. In W. Zhu, D. Tao, & X. Cheng (Eds.), *ACM Digital Library, Cikm'19: Proceedings of the 28th ACM International Conference on Information & Knowledge Management* (pp. 1261–1270). Association for Computing Machinery. https://doi.org/10.1145/3357384.3357995

Iglesias, A., Martínez, P., & Fernández, F. (1995). Navigating through the RLATES interface: A web-based adaptive and intelligent educational system. In G. Goos (Ed.), *Otm 2003 Workshops: Otm Confederated International Workshops, HCI-SWWA, IPW, JTRES,WORM, WMS, and WRSM 2003, Catania, Sicily, Italy, November 3–7, 2003. Proceedings* (1st ed.). Springer Berlin Heidelberg.

Iglesias, A., Martínez, P., & Fernández, F. (2003). An experience applying reinforcement learning in a web-based adaptive and intelligent educational system. *Informatics in Education*, *2*(2), 223–240. https://doi.org/10.15388/infedu.2003.17

Iglesias, A., Martínez, P., Aler, R., & Fernández, F. (2009a). Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Applied Intelligence, 31*(1), 89–106. https://doi.org/10.1007/s10489-008-0115-1

Iglesias, A., Martínez, P., Aler, R., & Fernández, F. (2009b). Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowledge-Based Systems, 22*(4), 266–270. https://doi.org/10.1016/j.knosys.2009.01.007

*Intayoad, W., Kamyod, C., & Temdee, P. (2018). Reinforcement learning for online learning recommendation system. In *The 6th Global Wireless Summit (GWS-2018): November 25–28, 2018, Mae Fah Luang University, Chiang Rai* (pp. 167–170). IEEE. https://doi.org/10.1109/GWS.2018.8686513

Intayoad, W., Kamyod, C., & Temdee, P. (2020). Reinforcement learning based on contextual bandits for personalized online learning recommendation systems. *Wireless Personal Communications, 115*(4), 2917–2932. https://doi.org/10.1007/s11277-020-07199-0

Islam, R., Henderson, P., Gomrokchi, M., & Precup, D. (2017). Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. In *ICML Reproducibility in Machine Learning Workshop, ICML'17*.

Islam, M. Z., Ali, R., Haider, A., Islam, M. Z., & Kim, H. S. (2021). PAKES: A reinforcement learning-based personalized adaptability knowledge extraction strategy for adaptive learning systems. *IEEE Access, 9*, 155123–155137. https://doi.org/10.1109/ACCESS.2021.3128578

JASP Team. (2021). *JASP* (Version 0.16.0.0) [Computer software]. https://jasp-stats.org/

Jatzlau, S., Michaeli, T., Seegerer, S., & Romeike, R. (2019). It's not magic after all: Machine learning in Snap! using reinforcement learning. In *2019 IEEE Blocks and Beyond Workshop (B & B): B & B 2019 : October 18, 2019, Memphis, Tennessee, USA : Proceedings* (pp. 37–41). IEEE. https://doi.org/10.1109/BB48857.2019.8941208

*Jeewantha, H. C. R., Gajasinghe, A. N., Naidabadu, N. I., Rajapaksha, T. N., Kasthurirathna, D., & Karunasena, A. (2021). English language trainer for non-native speakers using audio signal processing, reinforcement learning, and deep learning. In *21st International Conference on Advances in ICT for Emerging Regions (ICTer) 2021: Conference proceedings : 02nd & 03rd of December 2021, University of Colombo, School of Computing, Colombo, Sri Lanka* (pp. 117–122). IEEE. https://doi.org/10.1109/ICter53630.2021.9774785

Johnson, S., & Zaiane, O. R. (2013). Intelligent feedback polarity and timing selection in the Shufti intelligent tutoring system. In *International Conference on Computers in Education*.

*Ju, S., Zhou, G., Abdelshiheed, M., Barnes, T., spsampsps Chi, M. (2021). Evaluating critical reinforcement learning framework in the field. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, spsampsps V. Dimitrova (Eds.), *Lecture Notes in Computer Science. ARTIFICIAL INTELLIGENCE IN EDUCATION: 22nd international conference, aied 2021* (Vol. 12748, pp. 215–227). SPRINGER NATURE. https://doi.org/10.1007/978-3-030-78292-4_18

*Jung, G., Ausin, M. S., Barnes, T., & Chi, M. (2024). More, May not the Better: Insights from Applying Deep Reinforcement Learning for Pedagogical Policy Induction. In *Proceedings of the 17th International Conference on Educational Data Mining*.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence, 101*(1–2), 99–134. https://doi.org/10.1016/S0004-3702(98)00023-X

Kakdas, Y. C., Kockara, S., Halic, T., & Demirel, D. (2024). Enhancing Medical Training Through Learning From Mistakes by Interacting With an Ill-Trained Reinforcement Learning Agent. *IEEE Transactions on Learning Technologies, 17*, 1248–1260. https://doi.org/10.1109/tlt.2024.3372508

Kandel, A., Ibrahim, I., & Fukuta, N. (2022). An analysis of educational cloud platforms using multi-agent learning. In T. Matsui, K. Takamatsu, & Y. Ono (Eds.), *2022 12th International Congress on Advanced Applied Informatics: Iiai-AAI 2022 : Kanazawa, Japan, 2–7 July 2022 : Proceedings* (pp. 230–233). IEEE. https://doi.org/10.1109/IIAIAAI55812.2022.00053

*Kim, S., Kim, W., spsampsps Kim, H. (2021). Learning path construction using reinforcement learning and bloom's taxonomy. In A. I. Cristea spsampsps C. Troussas (Eds.), *Springer eBook Collection: Vol. 12677. Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings* (1st ed. 2021, Vol. 12677, pp. 267–278). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-80421-3_29

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., spsampsps Pineau, J. (2020). Automated personalized feedback improves learning gains in an intelligent tutoring system. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, spsampsps E. Millán (Eds.), *Springer eBook Collection: Vol. 12164. Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II* (1st ed. 2020, Vol. 12164, pp. 140–146). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-52240-7_26

Konda, V., & Tsitsiklis, J. (1999). Actor-critic algorithms. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in Neural Information Processing Systems* (Vol. 12). MIT Press.

Koroveshi, J., & Ktona, A. (2020). Modelling an intelligent tutoring system using reinforcement learning. *Knowledge - International Journal, 43*(3), 483–487.

Koroveshi, J., & Ktona, A. (2021). Training an intelligent tutoring system using reinforcement learning. *International Journal of Computer Science and Information Security (IJCSIS), 19*(3), 10–18. https://doi.org/10.5281/zenodo.4661454

Kubotani, Y., Fukuhara, Y., & Morishima, S. (2021). RLTutor: Reinforcement learning based adaptive tutoring system by modeling virtual student with fewer interactions. In *AI4EDU workshop at IJCAI2021*.

Kumar, A., spsampsps Ahuja, N. J. (2020). An adaptive framework of learner model using learner characteristics for intelligent tutoring systems. In S. Choudhury, R. Mishra, R. G. Mishra, spsampsps A. Kumar (Eds.), *Springer eBooks Intelligent Technologies and Robotics: Vol. 989. Intelligent*

Communication, Control and Devices: Proceedings of ICICCD 2018 (1st ed. 2020, Vol. 989, pp. 425–433). Springer. https://doi.org/10.1007/978-981-13-8618-3_45

Lagoudakis, M. G., & Parr, R. (2003). Least-squares policy iteration. *The Journal of Machine Learning Research*, (4), 1107–1149.

Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. *Cambridge University Press*. https://doi.org/10.1017/9781108571401

Leclercq, V., Beaudart, C., Ajamieh, S., Rabenda, V., Tirelli, E., & Bruyère, O. (2019). Meta-analyses indexed in PsycINFO had a better completeness of reporting when they mention PRISMA. *Journal of Clinical Epidemiology, 115*, 46–54. https://doi.org/10.1016/j.jclinepi.2019.06.014

Lee, J. in, & Brunskill, E. (2012). The impact on individualizing student models on necessary practice opportunities. In *International Conference on Educational Data Mining (EDM),* Chania, Greece.

Legaspi, R. S., & Sison, R. C. (2002). A machine learning framework for an expert tutor construction. In *Proceedings / International Conference on Computers in Education: December 3 - 6, 2002, Auckland, New Zealand* (pp. 670–674). IEEEE Computer Society. https://doi.org/10.1109/CIE.2002.1186038

Leite, W. L., Roy, S., Chakraborty, N., Michailidis, G., Huggins-Manley, A. C., D'Mello, S., Shirani Faradonbeh, M. K., Jensen, E., Kuang, H., & Jing, Z. (2022). A novel video recommendation system for algebra: An effectiveness evaluation study. In A. F. Wise (Ed.), *ACM Digital Library, Lak22: 12th International Learning Analytics and Knowledge Conference* (pp. 294–303). Association for Computing Machinery. https://doi.org/10.1145/3506860.3506906

Lenhard, W., & Lenhard, A. (2017). *Computation of effect sizes*. Psychometrica. https://doi.org/10.13140/RG.2.2.17823.92329

Li, Q., Xia, W., Yin, L., Jin, J., & Yu, Y. (2024a). Privileged Knowledge State Distillation for Reinforcement Learning-based Educational Path Recommendation. In R. Baeza-Yates & F. Bonchi (Eds.), *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1621–1630). ACM. https://doi.org/10.1145/3637528.3671872

Li, J., Yu, S., & Zhang, T. (2024b). Learning Path Recommendation Based on Reinforcement Learning. *Engineering Letters, 32*(9), 1823–1832.

Li, K. (2019). MOOC learners' demographics, self-regulated learning strategy, perceived learning and satisfaction: A structural equation modeling approach. *Computers & Education, 132*, 16–30. https://doi.org/10.1016/j.compedu.2019.01.003

Li, X., Xu, H., Zhang, J., & Chang, H.-H. (2023b). Deep reinforcement learning for adaptive learning systems. *Journal of Educational and Behavioral Statistics, 48*(2), 220–243. https://doi.org/10.3102/10769986221129847

Li, Z., Shi, L., Wang, J., Cristea, A. I., & Zhou, Y. (2023a). Sim-GAIL: A generative adversarial imitation learning approach of student modelling for intelligent tutoring systems. *Neural Computing and Applications, 35*(34), 24369–24388. https://doi.org/10.1007/s00521-023-08989-w

*Liang, K., & You, J. (2024). Research on personalized learning path recommendation model of artificial intelligence in new business. In *2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM)* (pp. 801–806). IEEE. https://doi.org/10.1109/ISPCEM64498.2024.00143

Lin, J., Ma, Z., Gomez, R., Nakamura, K., He, B., & Li, G. (2020). A review on interactive reinforcement learning from human social feedback. *IEEE Access, 8*, 120757–120765. https://doi.org/10.1109/ACCESS.2020.3006254

*Liu, S., Chen, Y., Huang, H., Xiao, L., & Hei, X. (2018). Towards smart educational recommendations with reinforcement learning in classroom. In M. J. W. Lee (Ed.), *Proceedings of 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE2018)* (pp. 1079–1084). IEEE. https://doi.org/10.1109/TALE.2018.8615217

*Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., & Wang, S. (2019). Exploiting cognitive structure for adaptive learning. In A. Teredesai (Ed.), *ACM Digital Library, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 627–635). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330922

Liu, Y., Tang, W., & Pareek, P. K. (2022). The dynamic mode construction of mixed english learning based on reinforcement learning. In *2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC-2022)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICMNWC56175.2022.10031831

Liu, Y., & Zoghi, B. (2023). Enhancing STEM Education using Machine Learning and Reinforcement Learning Techniques for Educational Software and Serious Games. In L. Gómez Chova, A. López

Martínez, & I. Candel Torres (Eds.), *EDULEARN Proceedings, EDULEARN23 Proceedings* (pp. 7148–7152). IATED. https://doi.org/10.21125/edulearn.2023.1871

Liu, F., Hu, X., Liu, S., Bu, C., & Wu, L. (2023). Meta multi-agent exercise recommendation: A game application perspective. In A. Singh (Ed.), *ACM Digital Library, Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 1441–1452). Association for Computing Machinery. https://doi.org/10.1145/3580305.3599429

Loh, H., Shin, D., Lee, S., Baek, J., Hwang, C., Lee, Y., Cha, Y., Kwon, S., Park, J., & Choi, Y. (2021). Recommendation for effective standardized exam preparation. In M. Scheffel (Ed.), *ACM Digital Library, Lak21: 11th International Learning Analytics and Knowledge Conference* (pp. 397–404). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448177

López-López, J. A., Page, M. J., Lipsey, M. W., & Higgins, J. P. T. (2018). Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Research Synthesis Methods*, *9*(3). https://doi.org/10.1002/jrsm.1310

Maclellan, C., & Gupta, A. (2021). Learning expert models for educationally relevant tasks using reinforcement learning. In *International Conference on Educational Data Mining (EDM).*

Madani, Y., Ezzikouri, H., Erritali, M., & Hssina, B. (2020). Finding optimal pedagogical content in an adaptive e-learning platform using a new recommendation approach and reinforcement learning. *Journal of Ambient Intelligence and Humanized Computing, 11*(10), 3921–3936. https://doi.org/10.1007/s12652-019-01627-1

Maier, M., VanderWeele, T. J., & Mathur, M. B. (2022). Using selection models to assess sensitivity to publication bias: A tutorial and call for more routine use. *Campbell Systematic Reviews, 18*(3), Article e1256. https://doi.org/10.1002/cl2.1256

*Malpani, A., Ravindran, B., & Murthy, H. (2011). Personalized intelligent tutoring system using reinforcement learning. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, May 18–20, 2011, Palm Beach, Florida, USA*.

*Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., & Popovic, Z. (2014). Offline policy evaluation across representations with applications to educational games. In *AAMAS '14, Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 1077–1084). International Foundation for Autonomous Agents and Multiagent Systems.

Mandel, T. S. (2017). *Better education through improved reinforcement learning* [Doctoral dissertation, University of Washington]. ProQuest Dissertations & Theses.

*Martin, K. N., spsampsps Arroyo, I. (2004). Agentx: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In J. C. Lester, R. M. Vicari, spsampsps F. Paraguaçu (Eds.), *Lecture Notes in Computer Science: Vol. 3220. Intelligent Tutoring Systems: 7th International Conference, ITS 2004 Proceedings* (Vol. 3220, pp. 564–572). Springer. https://doi.org/10.1007/978-3-540-30139-4_53

Mavridis, D., & Salanti, G. (2013). A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research, 22*(2), 133–158. https://doi.org/10.1177/0962280211432219

Mazon, C., Clément, B., Roy, D., Oudeyer, P.-Y., & Sauzéon, H. (2023). Pilot study of an intervention based on an intelligent tutoring system (ITS) for instructing mathematical skills of students with ASD and/or ID. *Education and Information Technologies, 28*(8), 9325–9354. https://doi.org/10.1007/s10639-022-11129-x

Memarian, B., & Doleck, T. (2024). A scoping review of reinforcement learning in education. *Computers and Education Open, 6*, Article 100175. https://doi.org/10.1016/j.caeo.2024.100175

*Ming, G. F., & Hua, S. (2010). Course-scheduling algorithm of option-based hierarchical reinforcement learning. In Z. Hu (Ed.), *2010 Second International Workshop on Education Technology and Computer Science: Etcs 2010 ; Wuhan, China, 6 - 7 March 2010 ; [proceedings* (pp. 288–291). IEEE. https://doi.org/10.1109/ETCS.2010.584

Minoofam, S. A. H., Bastanfard, A., & Keyvanpour, M. R. (2022). Ralf: An adaptive reinforcement learning framework for teaching dyslexic students. *Multimedia Tools and Applications, 81*(5), 6389–6412. https://doi.org/10.1007/s11042-021-11806-y

Mirea, A.-M., & Preda, M. C. (2009). Adaptive learning based on exercises fitness degree. In R. Baeza-Yates & P. Boldi (Eds.), *2009 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT): Wi-IAT 2009* (pp. 215–218). IEEE. https://doi.org/10.1109/WI-IAT.2009.266

Mishima, C., & Asada, M. (1999). Active learning from cross perceptual aliasing caused by direct teaching. In *Human and environment friendly robots with high intelligence and emotional quotients:*

Proceedings (pp. 1420–1425). IEEE Operations Center. https://doi.org/10.1109/IROS.1999.811678

Mitchell, C. M., Boyer, K. E., spsampsps Lester, J. C. (2013). A markov decision process model of tutorial intervention in task-oriented dialogue. In H. C. Lane, K. Yacef, J. Mostow, spsampsps A. Graesser (Eds.), *Lecture notes in computer science Lecture notes in artificial intelligence: Vol. 7926. Artificial intelligence in education: 16th international conference, AIED 2013 proceedings* (Vol. 7926, pp. 828–831). Springer. https://doi.org/10.1007/978-3-642-39112-5_123

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533. https://doi.org/10.1038/nature14236

Moerland, T. M., Broekens, J., Plaat, A., & Jonker, C. M. (2023). Model-based reinforcement learning: A survey. *Foundations and Trends® in Machine Learning*, *16*(1), 1–118. https://doi.org/10.1561/2200000086

*Mohana, R., Sekhar, K. C., Sen Gupta, S., Punithaasree, K. S., Dorcas E, G., & Muthuperumal, S. (2024). Increasing Learner Engagement in English Language Acquisition Through AI-Powered Gamification. In *2024 International Conference on Artificial Intelligence and Quantum Computation-Based Sensor Application (ICAIQSA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICAIQSA64000.2024.10882349

Monahan, G. E. (1982). State of the Art—A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms. *Management Science, 28*(1), 1–16. https://doi.org/10.1287/mnsc.28.1.1

Mu, T., Wang, S., Andersen, E., & Brunskill, E. (2018). Combining adaptivity with progression ordering for intelligent tutoring systems. In S. Klemmer (Ed.), *ACM Proceedings of the Fifth Annual ACM Conference on Learning at Scale* (pp. 1–4). ACM. https://doi.org/10.1145/3231644.3231672

Mu, T., Wang, S., Andersen, E., spsampsps Brunskill, E. (2021). Automatic adaptive sequencing in a webgame. In A. I. Cristea spsampsps C. Troussas (Eds.), *Springer eBook Collection: Vol. 12677. Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings* (1st ed. 2021, Vol. 12677, pp. 430–438). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-80421-3_47

Murphy, K. (2025). *Reinforcement learning: An overview*. http://arxiv.org/pdf/2412.05265

Mustapha, R., Soukaina, G., Mohammed, Q., & Es-Sâadia, A. (2023). Towards an adaptive e-learning system based on deep learner profile, machine learning approach, and reinforcement learning. *International Journal of Advanced Computer Science and Applications*, *14*(5). https://doi.org/10.14569/IJACSA.2023.0140528

Nadelson, L. S., & Seifert, A. L. (2017). Integrated STEM defined: Contexts, challenges, and the future. *The Journal of Educational Research, 110*(3), 221–223. https://doi.org/10.1080/00220671.2017.1289775

Nguyen, H., & La, H. (2019). Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)* (pp. 590–595). IEEE. https://doi.org/10.1109/IRC.2019.00120

Nie, A., Reuel, A.-K., spsampsps Brunskill, E. (2023). Understanding the impact of reinforcement learning personalization on subgroups of students in math tutoring. In N. Wang, G. Rebolledo-Mendez, V. Dimitrova, N. Matsuda, spsampsps O. C. Santos (Eds.), *Communications in Computer and Information Science: Vol. 1831. Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky: 24th International Conference, AIED 2023 Proceedings* (1st ed. 2023, Vol. 1831, pp. 688–694). Springer Nature Switzerland; Imprint Springer. https://doi.org/10.1007/978-3-031-36336-8_106

Nisansala, P., & Morawaka, A. (2019). Athwel: Gamification supportive tool for special educational centers in Sri Lanka. In *2019 IEEE 14th International Conference on Industrial and Information Systems: (ICIIS) : 18th-20th December, 2019 : Conference proceedings* (pp. 446–451). IEEE. https://doi.org/10.1109/ICIIS47346.2019.9063274

*Niu, S., & Cao, S. (2022). Get a sense of accomplishment in doing exercises: A reinforcement learning perspective. In *2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD)* (pp. 299–304). IEEE. https://doi.org/10.1109/CSCWD54268.2022.9776133

Nwana, H. (1990). Intelligent tutoring systems: an overview. *Artificial Intelligence Review*, *4*(4). https://doi.org/10.1007/BF00168958

*Oralbayeva, N., Shakerimov, A., Sarmonov, S., Kantoreyeva, K., Dadebayeva, F., Serkali, N., & Sandygulova, A. (2022). K-Qbot: Language learning chatbot based on reinforcement learning. In S. Šabanović (Ed.), *Hri '22: Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 963–967). IEEE. https://doi.org/10.1109/HRI53351.2022.9889428

*Orsoni, M., Pögelt, A., Duong-Trung, N., Benassi, M., Kravcik, M., spsampsps Grüttmüller, M. (2023). Recommending mathematical tasks based on reinforcement learning and item response theory. In C. Frasson, P. Mylonas, spsampsps C. Troussas (Eds.), *Lecture Notes in Computer Science: Vol. 13891. Augmented Intelligence and Intelligent Tutoring Systems: 19th International Conference, ITS 2023, Corfu, Greece, June 2–5, 2023, Proceedings* (1st ed. 2023, Vol. 13891, pp. 16–28). Springer Nature Switzerland; Imprint Springer. https://doi.org/10.1007/978-3-031-32883-1_2

*Oyuga Anne, D., & Maina, E. (2021). Reinforcement learning approach for adaptive e-learning based on multiple learner characteristics. *Open Journal for Information Technology*, *4*(2), 55–76. https://doi.org/10.32591/coas.ojit.0402.03055o

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & McKenzie, J. E. (2021). Prisma 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. *BMJ (Clinical Research Ed.), 372*, Article n160. https://doi.org/10.1136/bmj.n160

*Pan, J., & Yang, N. (2024). Application of Reinforcement Learning Algorithm in Personalized Music Teaching. In *2024 2nd International Conference on Mechatronics, IoT and Industrial Informatics (ICMIII)* (pp. 599–604). IEEE. https://doi.org/10.1109/ICMIII62623.2024.00118

Panic, N., Leoncini, E., de Belvis, G., Ricciardi, W., & Boccia, S. (2013). Evaluation of the endorsement of the preferred reporting items for systematic reviews and meta-analysis (PRISMA) statement on the quality of published systematic review and meta-analyses. *PLoS ONE, 8*(12), Article e83138. https://doi.org/10.1371/journal.pone.0083138

Papadimitriou, C. H., & Tsitsiklis, J. N. (1987). The Complexity of Markov Decision Processes. *Mathematics of Operations Research*, *12*(3), 441–450. http://www.jstor.org/stable/3689975

Park, H. W., Grover, I., Spaulding, S., Gomez, L., & Breazeal, C. (2019). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*, 687–694. https://doi.org/10.1609/aaai.v33i01.3301687

Patel, M., & Sajja, P. S. (2021). Application for multi-agent system: A case of customised elearning. In S. L. Chavan (Ed.), *2021 International Conference on Computing, Communication and Green Engineering (CCGE2021)* (pp. 1–6). IEEE. https://doi.org/10.1109/CCGE50943.2021.9776390

Perez, J., Dapena, E., Aguilar, J., & Carrillo, G. (2022). Reinforcement learning for estimating student proficiency in math word problems. In *2022 XVII Latin American Conference on Learning Technologies (LACLO)* (pp. 1–6). IEEE. https://doi.org/10.1109/LACLO56648.2022.10013399

Pérez, J., Dapena, E., & Aguilar, J. (2024). Emotions as implicit feedback for adapting difficulty in tutoring systems based on reinforcement learning. *Education and Information Technologies, 29*(16), 21015–21043. https://doi.org/10.1007/s10639-024-12699-8

*Pietquin, O., Daubigney, L., & Geist, M. (2011). Optimization of a tutoring system from a fixed set of data. In *SLaTE 2011* (pp. 1–4).

*Pögelt, A., Ihsberner, K., Pengel, N., Kravcik, M., Grüttmüller, M., spsampsps Hardt, W. (2024). Individualised Mathematical Task Recommendations Through Intended Learning Outcomes and Reinforcement Learning. In A. Sifaleras spsampsps F. Lin (Eds.), *Lecture Notes in Computer Science: Vol. 14798. Generative Intelligence and Intelligent Tutoring Systems* (Vol. 14798, pp. 117–130). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63028-6_10

Priya, S. S., Subhashini, R., & Akilandeswari, J. (2012). Learning agent based knowledge management in intelligent tutoring system. In *2012 International Conference on Computer Communication and Informatics (ICCCI 2012)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICCCI.2012.6158828

*Pu, Y., Wang, C., & Wu, W. (2020). A deep reinforcement learning framework for instructional sequencing. In *2020 IEEE International Conference on Big Data (Big Data)* (pp. 5201–5208). IEEE. https://doi.org/10.1109/BigData50022.2020.9378463

Puterman, M. L. (2005). *Markov decision processes: Discrete stochastic dynamic programming*. *Wiley series in probability and mathematical statistics. Applied probability and statistics section*. Wiley-Interscience. https://doi.org/10.1002/9780470316887

*Raghuveer, V. R., Tripathy, B. K., Singh, T., & Khanna, S. (2014). Reinforcement learning approach towards effective content recommendation in MOOC environments. In *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)* (pp. 285–289). IEEE. https://doi.org/10.1109/MITE.2014.7020289

Ramachandran, A., & Scassellati, B. (2014). Adapting difficulty levels in personalized robot-child tutoring interactions. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*. https://www.aaai.org/ocs/index.php/WS/AAAIW14/paper/viewPaper/8736

*Ravari, P. B., Jen Lee, K., Law, E., & Kulic, D. (2021). Effects of an adaptive robot encouraging teamwork on students' learning. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)* (pp. 250–257). IEEE. https://doi.org/10.1109/RO-MAN50785.2021.9515354

*Reddy, S., Levine, S., & Dragan, A. (2017). Accelerating human learning with deep reinforcement learning. In *NIPS'17 Workshop: Teaching Machines, Robots, and Humans* (pp. 5–9).

Riedmann, A., & Lugrin, B. (2023). Towards an Adaptive Pedagogical Agent in a Reading Intervention Using Reinforcement Learning. In B. Lugrin, M. Latoschik, S. von Mammen, S. Kopp, F. Pécune, & C. Pelachaud (Eds.), *ACM Digital Library, Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (pp. 1–3). Association for Computing Machinery. https://doi.org/10.1145/3570945.3607320

*Riedmann, A., Götz, J., D'Eramo, C., spsampsps Lugrin, B. (2024). Uli-RL: A Real-World Deep Reinforcement Learning Pedagogical Agent for Children. In A. Hotho spsampsps S. Rudolph (Eds.), *Lecture Notes in Artificial Intelligence: Vol. 14992. Ki 2024: Advances in Artificial Intelligence: 47th German Conference on AI, Würzburg, Germany, September 25–27, 2024, Proceedings* (1st ed. 2024, Vol. 1410). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-70893-0_25

Riley, R. D. (2009). Multivariate meta-analysis: The effect of ignoring within-study correlation. *Journal of the Royal Statistical Society Series a: Statistics in Society, 172*(4), 789–811. https://doi.org/10.1111/j.1467-985X.2008.00593.x

Rojas-Barahona, L. M., spsampsps Cerisara, C. (2014). Bayesian inverse reinforcement learning for modeling conversational agents in a virtual environment. In D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, spsampsps A. Gelbukh (Eds.), *Lecture notes in computer science Theoretical computer science and general issues: Vol. 8403. Computational linguistics and intelligent text processing: 15th International Conference [on Intelligent Text Processing and Computational Linguistics], CICLing 2014 proceedings* (Vol. 8403, pp. 503–514). Springer. https://doi.org/10.1007/978-3-642-54906-9_41

*Rowe, J. P., & Lester, J. C. (2015). Improving student problem solving in narrative-centered learning environments: A modular reinforcement learning framework. In C. Conati, N. Heffernan, A. Mitrovic, & M. F. Verdejo (Eds.), *Lecture notes in computer science Lecture notes in artificial intelligence: Vol. 9112. Artificial intelligence in education: 17th international conference, AIED 2015 proceedings* (Vol. 9112, pp. 419–428). Springer. https://doi.org/10.1007/978-3-319-19773-9_42

Roy, S., Crick, C., Kieson, E., & Abramson, C. (2018). A reinforcement learning model for robots as teachers. In J.-J. Cabibihan (Ed.), *Ieee RO-MAN 2018: The 27th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 294–299). IEEE. https://doi.org/10.1109/ROMAN.2018.8525563

Ruan, S., Nie, A., Steenbergen, W., He, J., Zhang, J. Q., Guo, M., Liu, Y., Dang Nguyen, K., Wang, C. Y., Ying, R., Landay, J. A., & Brunskill, E. (2024). Reinforcement learning tutor better supported lower performers in a math task. *Machine Learning, 113*(5), 3023–3048. https://doi.org/10.1007/s10994-023-06423-9

Ruipérez-Valiente, J. A., Staubitz, T., Jenner, M., Halawa, S., Zhang, J., Despujol, I., Maldonado-Mahauad, J., Montoro, G., Peffer, M., Rohloff, T., Lane, J., Turro, C., Li, X., Pérez-Sanagustín, M., & Reich, J. (2022). Large scale analytics of global and regional MOOC providers: Differences in learners' demographics, preferences, and perceptions. *Computers & Education, 180*, Article 104426. https://doi.org/10.1016/j.compedu.2021.104426

Sarma, B. H. S., spsampsps Ravindran, B. (2007). Intelligent tutoring systems using reinforcement learning to teach autistic students. In A. Venkatesh, T. Gonzalves, A. Monk, spsampsps K. Buckner (Eds.), *IFIP International Federation for Information Processing: Vol. 241. Home informatics and telematics: Ict for the next billion: Proceeding of IFIP TC 9, WG 9.3 HOIT 2007 Conference* (Vol. 241, pp. 65–78). Springer. https://doi.org/10.1007/978-0-387-73697-6_5

*Sawyer, R., Rowe, J., spsampsps Lester, J. (2017). Balancing learning and engagement in game-based learning environments with multi-objective reinforcement learning. In E. André, R. Baker, X. Hu, M. M. T. Rodrigo, spsampsps B. Du Boulay (Eds.), *Lecture Notes in Computer Science: Vol. 10331. Artificial Intelligence in Education: 18th International Conference, AIED 2017 Proceedings* (Vol. 10331, pp. 323–334). Springer International Publishing. https://doi.org/10.1007/978-3-319-61425-0_27

Scarlatos, A., Smith, D., Woodhead, S., spsampsps Lan, A. (2024). Improving the Validity of Automatically Generated Feedback via Reinforcement Learning. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, spsampsps I. I. Bittencourt (Eds.), *Lecture Notes in Computer Science: Vol. 14829. Artificial Intelligence in Education* (Vol. 14829, pp. 280–294). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64302-6_20

Schmucker, R., Pachapurkar, N., Bala, S., Shah, M., spsampsps Mitchell, T. (2023). Learning to give useful hints: Assistance action evaluation and policy improvements. In O. Viberg, I. Jivet, P. J. Muñoz-Merino, M. Perifanou, spsampsps T. Papathoma (Eds.), *Lecture Notes in Computer Science: Vol. 14200. Responsive and Sustainable Educational Futures: 18th European Conference on Technology Enhanced Learning, EC-TEL 2023 Proceedings* (1st ed. 2023, Vol. 14200, pp. 383–398). Springer Nature Switzerland; Imprint Springer. https://doi.org/10.1007/978-3-031-42682-7_26

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In F. Bach & D. Blei (Eds.), *Proceedings of Machine Learning Research, Proceedings of the 32nd International Conference on Machine Learning* (pp. 1889–1897). PMLR. https://proceedings.mlr.press/v37/schulman15.html

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. https://arxiv.org/pdf/1707.06347.pdf

Shakya, A. K., Pillai, G., & Chakrabarty, S. (2023). Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications, 231*, Article 120495. https://doi.org/10.1016/j.eswa.2023.120495

Sharma, P., & Li, Q. (2024). Designing Simulated Students to Emulate Learner Activity Data in an Open-Ended Learning Environment. In *Proceedings of the 17th International Conference on Educational Data Mining*.

*Shawky, D., spsampsps Badawi, A. (2018). A reinforcement learning-based adaptive learning system. In A. E. Hassanien, M. F. Tolba, M. Elhoseny, spsampsps M. Mostafa (Eds.), *Advances in Intelligent Systems and Computing: Vol. 723. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2018)* (Vol. 723, pp. 221–231). Springer International Publishing. https://doi.org/10.1007/978-3-319-74690-6_22

Shawky, D., spsampsps Badawi, A. (2019). Towards a personalized learning experience using reinforcement learning. In A. E. Hassanien (Ed.), *Studies in Computational Intelligence: Volume 801. Machine Learning Paradigms: Theory and Application* (Vol. 801, pp. 169–187). Springer International Publishing. https://doi.org/10.1007/978-3-030-02357-7_8

*Shen, S., Ausin, M. S., Mostafavi, B., & Chi, M. (2018a). Improving learning & reducing time: A constrained action-based reinforcement learning approach. In T. Mitrovic (Ed.), *ACM Conferences, Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization* (pp. 43–51). ACM. https://doi.org/10.1145/3209219.3209232

*Shen, S., Mostafavi, B., Lynch, C., Barnes, T., spsampsps Chi, M. (2018c). Empirically evaluating the effectiveness of pomdp vs. mdp towards the pedagogical strategies induction. In C. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, spsampsps B. Du Boulay (Eds.), *Vol. 10948. Artificial Intelligence in Education: 19th International Conference, AIED 2018 Proceedings, Part II* (Vol. 10948, pp. 327–331). Springer. https://doi.org/10.1007/978-3-319-93846-2_61

*Shen, D., Truong, T., & Weintz, C. (2021). Using q-learning to personalize pedagogical policies for addition problems. In *2021 International Conference on Signal Processing and Machine Learning: Conf-SPML 2021 Proceedings* (pp. 186–189). IEEE. https://doi.org/10.1109/CONF-SPML54095.2021.00043

Shen, S., Mostafavi, B., Barnes, T., & Chi, M. (2018b). Exploring induced pedagogical strategies through a markov decision process framework: Lessons learned. *Journal of Educational Data Mining, 10*(3), 27–68. https://doi.org/10.5281/ZENODO.3554713

Shin, J., & Bulut, O. (2022). Building an intelligent recommendation system for personalized test scheduling in computerized assessments: A reinforcement learning approach. *Behavior Research Methods, 54*(1), 216–232. https://doi.org/10.3758/s13428-021-01602-9

Simmonds, M. (2015). Quantifying the risk of error when interpreting funnel plots. *Systematic Reviews, 4*, 24. https://doi.org/10.1186/s13643-015-0004-8

Singla, A., Rafferty, A. N., Radanovic, G., & Heffernan, N. T. (2021). Reinforcement learning for education: Opportunities and challenges [Workshop]. In *International Conference on Educational Data Mining (EDM)*.

Soto Forero, D., Ackermann, S., Laure Betbeder, M., & Henriet, J. (2024). Automatic Real-Time Adaptation of Training Session Difficulty Using Rules and Reinforcement Learning in the AI-VT ITS. *International Journal of Modern Education and Computer Science, 16*(3), 56–71. https://doi.org/10.5815/ijmecs.2024.03.05

*Spain, R., Rowe, J., Smith, A., Goldberg, B., Pokorny, R., Mott, B., & Lester, J. (2021). A reinforcement learning approach to adaptive remediation in online training. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 154851292110283. https://doi.org/10.1177/15485129211028317

*Stamper, J., Barnes, T., Lehmann, L., & Croy, M. (2008). The hint factory: Automatic generation of contextualized help for existing computer aided instruction. In *Proceedings of the 9th International Conference on Intelligent Tutoring Systems Young Researchers Track* (pp. 71–78).

Stanley, T. D. (2017). Limitations of PET-PEESE and Other Meta-Analysis Methods. *Social Psychological and Personality Science, 8*(5), 581–591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60–78. https://doi.org/10.1002/jrsm.1095

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Systematic reviews in health care: Investigating and dealing with publication and other biases in meta-analysis. *BMJ : British Medical Journal, 323*(7304), 101–105. https://doi.org/10.1136/bmj.323.7304.101

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ (Clinical Research Ed.), 343*, Article d4002. https://doi.org/10.1136/bmj.d4002

*Su, P.-h., Wang, Y.-B., Yu, T.-h., & Lee, L.-s. (2013). A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning. In *Icassp 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8213–8217). IEEE. https://doi.org/10.1109/ICASSP.2013.6639266

*Sun, M., Li, P., & Wang, D. (2024). Simulation and Optimization of Physical Education Teaching Based on Virtual Reality Technology and Reinforcement Learning Algorithms. In *2024 International Conference on Telecommunications and Power Electronics (TELEPE)* (pp. 579–584). IEEE. https://doi.org/10.1109/TELEPE64216.2024.00110

Sutton, R. S., & Barto, A. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Tang, Y., Hare, R., & Ferguson, S. (2022). Classroom evaluation of a gamified adaptive tutoring system. In *Fie 2022 Proceedings* (pp. 1–5). IEEE. https://doi.org/10.1109/FIE56618.2022.9962718

Tang, X., Chen, Y., Li, X., Liu, J., & Ying, Z. (2019). A reinforcement learning approach to personalized learning recommendation systems. *The British Journal of Mathematical and Statistical Psychology, 72*(1), 108–135. https://doi.org/10.1111/bmsp.12144

Teixeira da Silva, J. A., & Daly, T. (2024). Against Over-reliance on PRISMA Guidelines for Meta-analytical Studies. *Rambam Maimonides Medical Journal*, *15*(1). https://doi.org/10.5041/RMMJ.10518

Tetreault, J. R., & Litman, D. J. (2006b). Comparing the utility of state features in spoken dialogue using reinforcement learning. In Moore, R. C. (Ed.), *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics* - (pp. 272–279). Association for Computational Linguistics. https://doi.org/10.3115/1220835.1220870

*Tetreault, J., & Litman, D. (2006a). Using reinforcement learning to build a better model of dialogue state. In *EACL 2006, 11st Conference of the European Chapter of the Association for Computational Linguistics Proceedings*.

Thede, S. M. (2002). Using reinforcement learning to introduce artificial intelligence in the CS curriculum. *Journal of Computing Sciences in Colleges, 18*(1), 107–112.

Vahidy, J. (2019). Enhancing STEM learning through technology. In R. Power (Ed.), *Technology and the curriculum: Summer 2019*. Power Learning Solution.

VanLehn, K., Jordan, P., & Litman, D. (2007). Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *SLaTE-2007* (pp. 17–20).

*Vassoyan, J., Vie, J.-J., & Lemberger, P. (2023). Towards Scalable Adaptive Learning with Graph Neural Networks and Reinforcement Learning. In *Proceedings of the 16th International Conference on Educational Data Mining*.

Velusamy, B., Anouneia, S. M., & Abraham, G. (2013). Reinforcement learning approach for adaptive e-learning systems using learning styles. *Information Technology Journal, 12*(12), 2306–2314. https://doi.org/10.3923/itj.2013.2306.2314

Vijayan, A., Janmasree, S., Keerthana, C., & Baby Syla, L. (2018, July 5–7). A framework for intelligent learning assistant platform based on cognitive computing for children with autism spectrum disorder. In *2018 International CET Conference on Control, Communication, and Computing (IC4)* (pp. 361–365). IEEE. https://doi.org/10.1109/CETIC4.2018.8530940

*Wan, H., Che, B., Luo, H., & Luo, X. (2023). Learning path recommendation based on knowledge tracing and reinforcement learning. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 55–57). IEEE. https://doi.org/10.1109/ICALT58122.2023.00021

*Wang, F. (2014b). Pomdp framework for building an intelligent tutoring system. In S. Zvacek (Ed.), *Proceedings of the 6th International Conference on Computer Supported Education, Barcelona, Spain, 1 - 3 April, 2014* (pp. 233–240). SCITEPRESS. https://doi.org/10.5220/0004801702330240

Wang, F. (2014a). Learning teaching in teaching: Online reinforcement learning for intelligent tutoring. In C.-h. Pak, I. Stojmenovic, M. Choi, spsampsps F. Xhafa (Eds.), *Lecture Notes in Electrical Engineering: Vol. 276. Future information technology: Futuretech 2013* (Vol. 276, pp. 191–196). Springer. https://doi.org/10.1007/978-3-642-40861-8_29

Wang, L., Zhang, D., Gao, L., Song, J., Guo, L., & Shen, H. T. (2018). MathDQN: Solving arithmetic word problems via deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.11981

Wang, F. (2018). Reinforcement learning in a pomdp based intelligent tutoring system for optimizing teaching strategies. *International Journal of Information and Education Technology*, *8*(8), 553–558. https://doi.org/10.18178/ijiet.2018.8.8.1098

Wang, Y., Cai, W., Chen, M., & Shen, J. (2020). Poem: A personalized online education scheme based on reinforcement learning. In H. Mitsuhara (Ed.), *Proceedings of 2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE): Date and venue: 8–11 December 2020, online* (pp. 474–481). IEEE. https://doi.org/10.1109/TALE48869.2020.9368369

*Wang, J., Zhang, Y., Sun, L., Liu, Y., Zhang, W., & Zhang, Y. (2023). Learning path design on knowledge graph by using reinforcement learning. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 3480–3485). IEEE. https://doi.org/10.1109/BIBM58861.2023.10386061

*Wang, W., & Song, S. (2024). Real-Time Wireless Adaptive Learning Systems Using Reinforcement Learning and IoT for Smart Education. In *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)* (pp. 1–4). IEEE. https://doi.org/10.1109/CSRSWTC64338.2024.10811643

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

Whitehill, J., & Movellan, J. (2018). Approximately optimal teaching of approximately optimal learners. *IEEE Transactions on Learning Technologies, 11*(2), 152–164. https://doi.org/10.1109/TLT.2017.2692761

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning, 8*(3–4), 229–256. https://doi.org/10.1007/BF00992696

Wu, S., Wang, J., & Zhang, W. (2024). Contrastive Personalized Exercise Recommendation With Reinforcement Learning. *IEEE Transactions on Learning Technologies, 17*, 691–703. https://doi.org/10.1109/TLT.2023.3326449

Yang, J. (2024). English Learning Knowledge Point Recommendation Algorithm based on Deep Deterministic Policy Gradient. In *2024 International Conference on Integrated Intelligence and Communication Systems (ICIICS)* (pp. 1–5). IEEE. https://doi.org/10.1109/ICIICS63763.2024.10860216

*Yantao, L., & Wei, J. (2024). Enhancing Student Engagement in Smart Classrooms Using Reinforcement Learning Algorithms. In *2024 Cross Strait Radio Science and Wireless Technology Conference (CSRSWTC)* (pp. 1–4). IEEE. https://doi.org/10.1109/CSRSWTC64338.2024.10811575

*Yessad, A. (2023). Using the ITS components in improving the q-learning policy for instructional sequencing. In C. Frasson, P. Mylonas, spsampsps C. Troussas (Eds.), *Lecture Notes in Computer Science: Vol. 13891. Augmented Intelligence and Intelligent Tutoring Systems: 19th International Conference, ITS 2023 Proceedings* (1st ed. 2023, Vol. 13891, pp. 247–256). Springer Nature Switzerland; Imprint Springer. https://doi.org/10.1007/978-3-031-32883-1_21

Yuh, M. S., Rabb, E., Thorpe, A., & Jain, N. (2024). Using Reward Shaping to Train Cognitive-Based Control Policies for Intelligent Tutoring Systems. In *2024 American Control Conference (ACC)* (pp. 3223–3230). IEEE. https://doi.org/10.23919/ACC60939.2024.10644169

Zadem, M., Mover, S., & Nguyen, S. M. (2023). Emergence of a Symbolic Goal Representation with an Intelligent Tutoring System based on Intrinsic Motivation. In *NeurIPS 2023: IMOL Workshop "Intrinsically-Motivated and Open-Ended Learning"* (pp. 423–428). IEEE.

*Zhang, Y., & Goh, W.-B. (2019). Bootstrapped policy gradient for difficulty adaptation in intelligent tutoring systems. In *AAMAS '19, Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 711–719). International Foundation for Autonomous Agents and Multiagent Systems.

Zhang, H., spsampsps Yu, T. (2020). Taxonomy of reinforcement learning algorithms. In H. Dong, Z. Ding, spsampsps S. Zhang (Eds.), *Springer eBook Collection. Deep Reinforcement Learning: Fundamentals, Research and Applications* (1st ed. 2020, pp. 125–133). Springer Singapore; Imprint Springer. https://doi.org/10.1007/978-981-15-4095-0_3

*Zhang, J. (2023). Game Design and Learning Effectiveness Evaluation of English Teaching Based on Reinforcement Learning Algorithm. In *2023 International Conference on Intelligent Computing, Communication & Convergence (ICI3C)* (pp. 349–353). IEEE. https://doi.org/10.1109/ICI3C60830.2023.00073

Zhang, D. (2024). Using deep Reinforcement Learning to Optimize the Motivational Incentive Mechanism of Online English Learners. In *Proceedings of the International Conference on Decision Science & Management* (pp. 179–183). ACM. https://doi.org/10.1145/3686081.3686110

Zhang, Y., & Goh, W.-B. (2021). Personalized task difficulty adaptation based on reinforcement learning. *User Modeling and User-Adapted Interaction, 31*(4), 753–784. https://doi.org/10.1007/s11257-021-09292-w

Zhiyong, J., Jing, T., & Jing, Z. (2021). Allocation of english remote guiding based on deep reinforcement learning and multi-objective optimization. In *Proceedings of the 5th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud): I-SMAC 2021 : 11–13, November 2021* (pp. 414–417). IEEE. https://doi.org/10.1109/I-SMAC52330.2021.9640763

*Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., spsampsps Chi, M. (2019). Hierarchical reinforcement learning for pedagogical policy induction. In S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, spsampsps R. Luckin (Eds.), *LNCS sublibrary: 11625–11626. Artificial intelligence in education: 20th international conference, AIED 2019, Chicago, IL, USA, June 25–29, 2019, proceedings* (Vol. 11625, pp. 544–556). Springer International Publishing. https://doi.org/10.1007/978-3-030-23204-7_45

*Zhou, G., Yang, X., Azizsoltani, H., Barnes, T., & Chi, M. (2020). Improving student-system interaction through data-driven explanations of hierarchical reinforcement learning induced pedagogical policies. In T. Kuflik (Ed.), *ACM Digital Library, Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 284–292). Association for Computing Machinery. https://doi.org/10.1145/3340631.3394848

Zhou, G., Azizsoltani, H., Ausin, M. S., Barnes, T., & Chi, M. (2021). Leveraging granularity: Hierarchical reinforcement learning for pedagogical policy induction. *International Journal of Artificial Intelligence in Education, 32*(2), 454–500. https://doi.org/10.1007/s40593-021-00269-9