informatics

MDPI

Article

Reinforcement Learning in Education: A Literature Review

Bisni Fahad Mon [1,†], Asma Wasfi [2,†], Mohammad Hayajneh [1,*], Ahmad Slim [3] and Najah Abu Ali [1]

[1] Computer and Network Engineering, College of Information Technology, UAE University, Al Ain 15551, United Arab Emirates; bisni.f@uaeu.ac.ae (B.F.M.); najah@uaeu.ac.ae (N.A.A.)
[2] Electrical and Communication Engineering, UAE University, Al Ain 15551, United Arab Emirates; 201180954@uaeu.ac.ae
[3] Electrical and Computer Engineering, The University of Arizona, Tucson, AZ 85721, USA; ahslim@arizona.edu
* Correspondence: mhayajneh@uaeu.ac.ae
† These authors contributed equally to this work.

Abstract: The utilization of reinforcement learning (RL) within the field of education holds the potential to bring about a significant shift in the way students approach and engage with learning and how teachers evaluate student progress. The use of RL in education allows for personalized and adaptive learning, where the difficulty level can be adjusted based on a student's performance. As a result, this could result in heightened levels of motivation and engagement among students. The aim of this article is to investigate the applications and techniques of RL in education and determine its potential impact on enhancing educational outcomes. It compares the various policies induced by RL with baselines and identifies four distinct RL techniques: the Markov decision process, partially observable Markov decision process, deep RL network, and Markov chain, as well as their application in education. The main focus of the article is to identify best practices for incorporating RL into educational settings to achieve effective and rewarding outcomes. To accomplish this, the article thoroughly examines the existing literature on using RL in education and its potential to advance educational technology. This work provides a thorough analysis of the various techniques and applications of RL in education to answer questions related to the effectiveness of RL in education and its future prospects. The findings of this study will provide researchers with a benchmark to compare the usefulness and effectiveness of commonly employed RL algorithms and provide direction for future research in education.

Keywords: education; artificial intelligence; reinforcement learning; Markov decision process

check for updates

Citation: Fahad Mon, B.; Wasfi, A.; Hayajneh, M.; Slim, A.; Abu Ali, N. Reinforcement Learning in Education: A Literature Review. *Informatics* **2023**, *10*, 74. https://doi.org/10.3390/informatics10030074

Academic Editor: Sotiris Kotsiantis

Received: 3 July 2023
Revised: 10 September 2023
Accepted: 11 September 2023
Published: 18 September 2023

1. Introduction

Artificial intelligence (AI) technologies are advancing rapidly and have a significant impact on modern society. As a result, AI has become a major topic in education among other fields [1–6]. The use of AI technologies, such as intelligent tutoring systems, chatbots, robots, and the automated evaluation of all forms of digital artifacts that support and enhance education, is known as AI in education (AIEd). While being a subcategory of machine learning (ML), RL is also a universal framework for AI and automated decision making. RL, supervised learning, and unsupervised learning are the three main types of ML. Models are pretrained using labeled datasets during supervised learning. Unsupervised learning, on the other hand, involves training the model on unlabeled datasets. Figure 1 represents the relationship between artificial intelligence (AI), machine learning, reinforcement learning, and deep learning. RL can be defined as a computational paradigm for sequential decision making and goal-directed learning [7–10]. Contrary to supervised learning and unsupervised learning, RL focuses on agent learning through direct contact with its environment. Consecutive decisions can be made because RL helps agents to efficiently interact with their environment. It encourages behavioral decision making by utilizing interaction experience and evaluating feedback afterward [7]. The agent receives a reward or evaluative feedback and a new state based on the action selected for the current state. Then, it learns the optimal path that will maximize its reward by acquiring experience and knowledge rather than by receiving instructions. RL is especially well suited for scenarios where an agent must develop a policy on what to do in various circumstances and how to map states to actions in order to enhance long-term utility. The agent must experiment with various actions to find those that produce high rewards. These actions not only affect the next state but also future rewards.

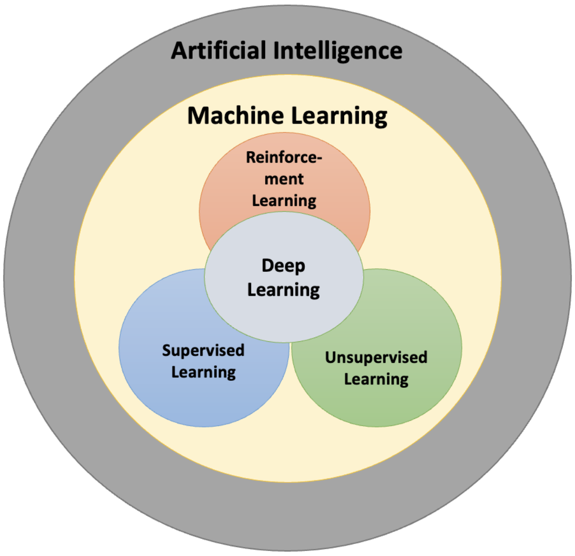

Figure 1. Schematic representation of the relationship between artificial intelligence (AI), machine learning, reinforcement learning, and deep learning.

Over the past 10 years, numerous RL algorithms and methods have been developed to address real-world issues. The Markov decision process (MDP), Monte Carlo, Q-learning, temporal difference, and dynamic programming are among them. RL has applications in several fields, including smart grids, engineering, healthcare, robotics, finance, transportation, and natural sciences. One of the interesting applications of RL has been observed in the gaming domain [11–16]. There has been an increase in interest in the goal-directed enhancement of state-of-the-art technology for education. Specifically, sequential student–teacher interactions are essential for effective learning. RL offers itself as a handy tool for a variety of problem contexts in education, including developing teaching strategies using RL techniques and simulating a human student using RL. The use of conventional RL approaches in education is challenging, despite their potential.

RL is commonly seen as an effective framework for enhancing educational outcomes by numerous researchers. As a growing number of researchers are exploring the use of reinforcement learning to advance education, it is crucial to consider and implement the best practices in this field. In this paper, we study RL's impact on education through an exploration of its applications and techniques within this context. Incorporating reinforcement learning into educational settings can lead to successful and rewarding outcomes.

In this paper, we proceed by surveying the various applications of reinforcement learning in education. The structure of the paper is as follows. Section 2 illustrates the methodology followed to conduct this survey. Section 3 reviews the RL techniques and applications in the education field. Considerations for the design of reinforcement learning are discussed in Section 4. In Section 5, the challenges of AI in education and future research directions are illustrated. Finally, Sections 6 and 7 discuss and conclude the work of this study, respectively.

2. Review Planning and Methodology

In this literature review article, a systematic review method was followed. This work involved summarizing and synthesizing the findings of the existing literature on the following research questions:

RQ1: Does RL actually help in the education field?
RQ2: If so, what are the applications, and where might we anticipate it being most useful?
RQ3: What are the considerations, challenges, and future directions of RL in education?

Figure 2 illustrates the methodology utilized during this study. Initially, we identified specific research questions that guided our literature review approach. Subsequently, we conducted an extensive search using respected academic databases such as Google Scholar, IEEE Explore, and Scopus- with the aim of retrieving relevant articles discussing reinforcement learning methods within educational domains while filtering out irrelevant sources via our screening process criteria. We used a combination of terms and phrases related to both reinforcement learning and education to ensure a thorough search for pertinent material. The terms "reinforcement learning", "education", "learning algorithms", and "educational technology" were selected as keywords. These keywords were chosen based on their applicability to our subject and their propensity to find a large number of relevant articles. Next, a screening process was conducted to include only those articles that focused on reinforcement learning in education. The selected articles were then analyzed and synthesized to answer the research questions. The PRISMA flow diagram with database searches, the number of abstracts examined, and the complete texts retrieved for the review are shown in Figure 3. During this process, gaps in the literature were identified, which could be addressed in future research. Finally, given these results, we synthesized specific information gathered from these sources, enabling us to address our specified inquiry effectively while noting key areas which require further investigation. RL has the potential to bring about a significant shift in the way students approach and engage with learning and how teachers evaluate students' progress. Our paper aims to answer important questions related to the effectiveness and potential of RL algorithms in the field of education. Specifically, we provide answers to crucial questions such as whether RL actually helps in the education field, what its applications are, and where we might anticipate it being most useful. We also explore the future direction of RL in education. To accomplish our objective, we surveyed the various applications of RL in education and compared the various policies induced by RL with baselines. We identified four distinct RL techniques: the Markov decision process, partially observable Markov decision process, deep RL network, and Markov chain as well as their application in education. We also identified the best practices for incorporating RL into educational settings to achieve effective and rewarding outcomes. Furthermore, we thoroughly examined the existing literature on the use of RL in education and its potential to advance educational technology.

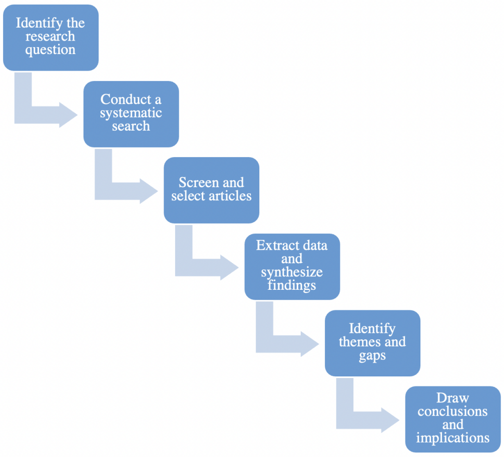
Identify the research question
Conduct a systematic search
Screen and select articles
Extract data and synthesize findings
Identify themes and gaps
Draw conclusions and implications

Figure 2. Research methodology.

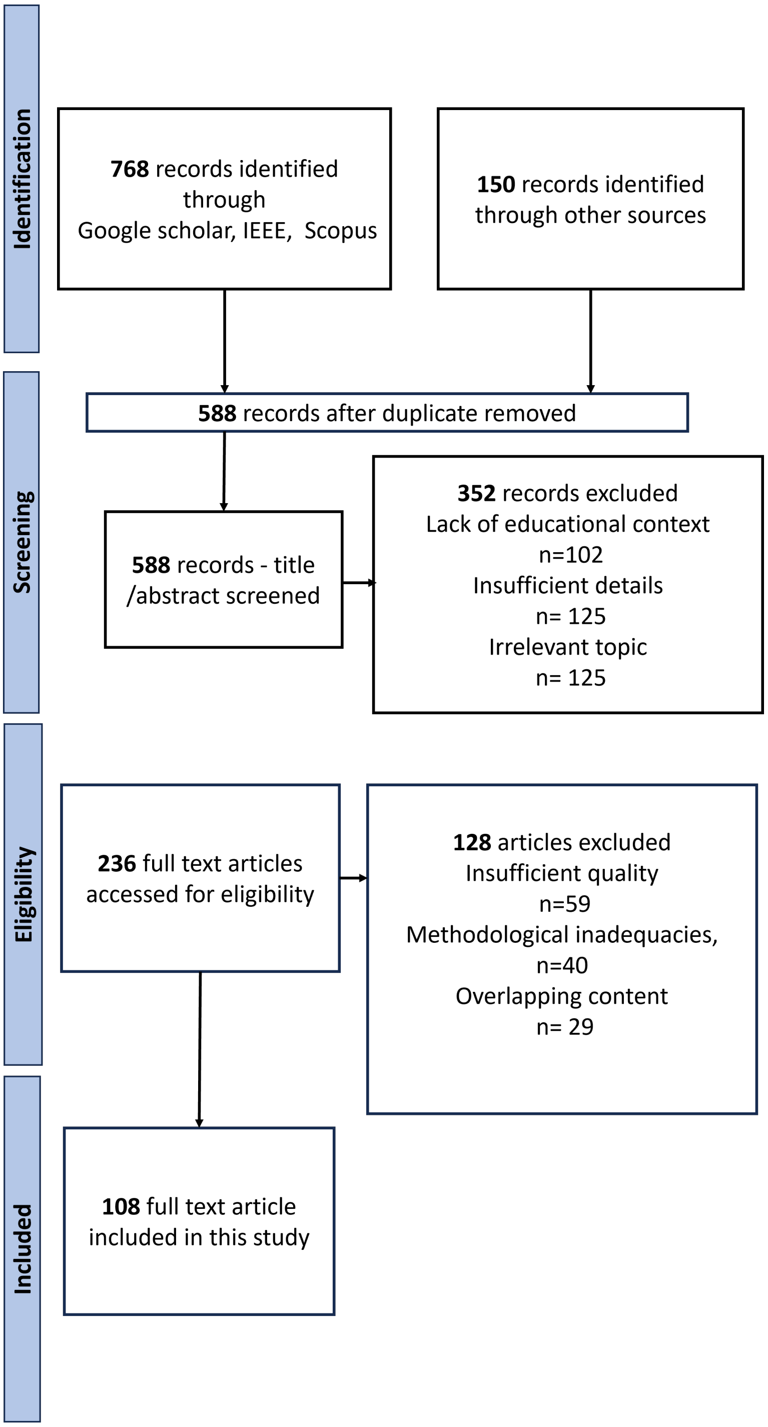
Identification
Screening
Eligibility
Included
768 records identified through Google scholar, IEEE, Scopus
150 records identified through other sources
588 records after duplicate removed
588 records - title /abstract screened
352 records excluded
Lack of educational context
n=102
Insufficient details
n= 125
Irrelevant topic
n= 125
236 full text articles accessed for eligibility
128 articles excluded
Insufficient quality
n=59
Methodological inadequacies,
n=40
Overlapping content
n= 29
108 full text article included in this study

Figure 3. The PRISMA flow diagram showing database searches, the number of abstracts examined, and the complete texts retrieved for systematic review.

3. Literature Review

This literature review article provides a comprehensive review of RL applications and techniques and their potential impact on education while also highlighting the best practices and future research directions. Researchers will be able to compare the usefulness and effectiveness of commonly employed RL algorithms in education, which will guide them in new directions. This section presents reviews on RL research directions in education application, RL techniques in the education domain, and the application of RL techniques in curricular analytics in higher education.

3.1. RL Techniques in the Educational Domain

RL algorithms and methods are being extensively used in the education domain to enhance student performance, facilitate the teacher tutoring process, reduce the time needed to acquire or gain knowledge, and improve the students' graduation rate. This section illustrates the various RL techniques used in different educational applications.

3.1.1. Markov Decision Process

The MDP provides decision makers with improved management and planning strategies by finding the best decision in a critical system such as educational application. The MDP is a decision support system that is used to model dynamics in order to solve problems related to finding the most profitable method, where it studies various decisions before choosing the best one. Dynamic programming algorithms can be used to identify the method that will give the best value for the investment in its various fields. Several targeting scenarios can be simulated to find out the expected impact of a policy [17].

MDPs have four main elements: A refers to a finite set of actions, S refers to a finite set of states, $P_a(s, s')$ refers to the probability that action a in state s at time t leads to state s' at time $t + 1$, and $R_a(s, s')$ refers to the reward acquired after transitioning from state s to state s' via action a. The agent is the decision-making entity within the MDP. It interacts with the environment by selecting actions from the finite set of actions based on the current state of the environment. Beck et al. [18] studied the learning of temporal differences (TDs) [19] to place educational policies that would reduce the time students spend to complete mathematical problems through AnimalWatch, an intelligent tutor used to teach arithmetic to school students. In the training phase of the study, simulated students were modeled as computer agents, and time was used as an immediate reward. In the test phase, the original version of AnimalWatch was compared to the new version of AnimalWatch with the educational policy. The authors observed that the group of students who used the new version of AnimalWatch spent less time on each mathematical problem compared with the students who used the older version. Thus, the new version with induced policy improved the students' performance. Iglesias et al. induced a policy through RLATES, an intelligent system that educates students on the design of the database. They used online Q-learning with time as the immediate reward [20–22]. The aim of the applied policy is to support students with direct navigation via the content of the system to enhance their learning process. Simulated students were utilized in the training phase. The induced policy performance was evaluated by comparing both real and simulated students using RLATES with students using IGNATES, which supports indirect navigation without reinforcement learning. It was noticed that the students who were using RLATES needed less time than the ones who were using IGNATES. However, both systems provided the same level of knowledge, which was assessed by conducting an exam. Martin and Arroyo used Wayang Outpost [23], a web-based intelligent tutoring system. The authors applied a model-based reinforcement learning technique with a delayed reward to generate policies to increase the efficiency of hint sequencing. A student model was used to produce the training data, which were used to produce the policies. The test phase included testing the new RL policies on simulated students and evaluating the students' performance. The results indicate that the newly induced RL policies led to a better learning level. Chi et al. [24] improved the efficiency of Cordillera, an intelligent physics tutoring system, by applying model-based RL. The research group gathered exploratory data by training students on the Cordillera version with random decisions. The system evaluation showed that the newly employed RL policies were more effective than the previous policies of normalized learning gain (NLG).

3.1.2. Partially Observable Markov Decision Process (POMDP)

The POMDP is another framework that is widely used in educational fields [25,26]. Mendel et al. [27] integrated the POMDP with a feature compression approach to induce policies for an educational game. It was noticed that the induced policies with immediate rewards had better results than expert-designed policies and random policies in both empirical and simulated evaluations. For example, the POMDP was applied by Rafferty et al. [28] to present students' latent knowledge. They applied the POMDP to induce policies that utilized time as a reward with the aim of reducing the time needed for learners to comprehend concepts. The POMDP-induced policies were evaluated by both real-world and simulated studies, and it was noticed that these policies outperformed the random policy. Clement et al. [29] developed a model to follow the students' knowledge of each component. They integrated student models with the POMDP to induce teaching policies while the immediate reward was the gained learning. The model was evaluated by using a series of simulated studies. The results showed that the POMDP policies had better performance than the learning theory-based policies. Whitehill and Movellan [17] employed the POMDP to minimize the required time for foreign language learning by inducing a teaching policy. They used a modified student model to construct the belief state of their POMDP. Furthermore, they conducted a real-world study to ensure that the POMDP policy performed better compared with two hand-crafted teaching policies.

3.1.3. Deep RL Framework

There is growing interest in the deep RL framework for inducing policies. Deep RL uses deep neural networks for the state approximation and function approximation [11,12]. This improvement enables the agent to handle challenging tasks. Wang et al. [30] used a deep RL framework in an educational game, CRYSTAL ISLAND, to personalize its interactive narratives. Immediate rewards were used based on normalized learning gain (NLG). Simulation studies showed that the students' NLG scores after applying the deep RL policy were higher than those who used linear RL.

3.1.4. Markov Chain

The Markov chain model teaching assessment method is a quantitative analytic technique based on stochastic process theory and probability theory. It creates a stochastic mathematical model to examine the quantitative relationship in the change and development process of actual activities. The use of it to evaluate college teachers' classroom teaching abilities in a more thorough, reasonable, and effective manner will help to advance the ongoing development of teachers' skill levels and the caliber of education provided in classrooms. This research presents an improved Markov chain hybrid teaching quality evaluation model after carefully examining the Markov chain algorithm theory, designing comparative tests, and then implementing it in the hybrid teaching quality evaluation system of universities. It then creates a comparable hybrid quality evaluation model of teaching and, in the end, conducts studies to confirm its effectiveness. This study's mathematical model of mixed-classroom teaching quality evaluation focuses on how the teaching process is changed and developed over time. In [31], a fuzzy assessment of the quality of physical education was performed, and examples were used to demonstrate its viability. The authors accomplished this by utilizing computer technology, the Markov chain, and the index system of the student's quality of sports. This kind of appraisal effort has good practical significance given the prevalence of information and computer technologies in physical education. The standard of physical education students and instructors is determined by a dynamic process in which a student who performed exceptionally well on one assessment may barely pass the next. Changes in the standard of PE were reflected in the results of a thorough analysis of the kids' transfer states. Here, the Markov chain provided a tool for dynamic analysis.

There are numerous events in the actual world that fit this description, provided that one is aware of the current state of a given system. One can infer its future state by looking at the present rather than at its past. This occurrence is referred to as the Markov random phenomenon. The Markov model is a mathematical representation of this random event [31]. The model developed in [32] is more accurate and reasonable for assessing the quality of teaching for the teaching process that is directly related to the attribution of teaching quality. Figure 4 illustrates the fundamental steps of blended learning using the Markov algorithm.

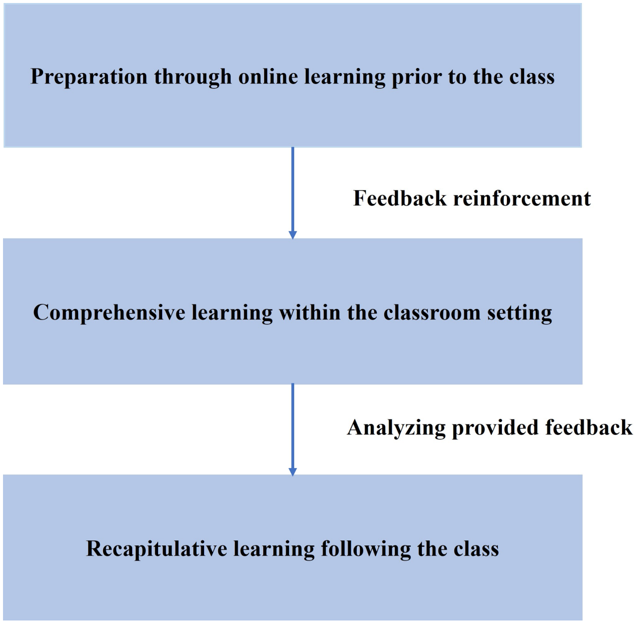
Preparation through online learning prior to the class
Feedback reinforcement
Comprehensive learning within the classroom setting
Analyzing provided feedback
Recapitulative learning following the class

Figure 4. Blended teaching fundamentals based on Markov algorithm.

3.2. RL Research Directions in the Education Application

RL algorithms and methods have been used extensively in various research directions in education applications, as explained in the following subsections.

3.2.1. RL Techniques for the Teacher–Student Framework

The teacher–student framework has been introduced as a way to improve sample efficiency by deploying an advising mechanism. This mechanism involves a teacher guiding the student's exploration. Previous studies in this field have focused on the teacher advising the student on the optimal action to take in a given state. However, Anand et al. [33] proposed extending this advising mechanism by incorporating qualitative assessments of the state provided by the teacher, leveraging their domain expertise to provide more informative signals. To effectively reuse the advice provided by the teacher, a novel architecture called Advice Replay Memory (ARM) has been introduced. The robustness of this approach has been demonstrated through experiments on multiple Atari 2600 games using a fixed set of hyperparameters. The results show that even with a suboptimal teacher, the student can still achieve significant performance boosts and eventually outperform the teacher. The proposed approach outperforms the baselines even when provided with a smaller advising budget and comparatively suboptimal teachers. Zimmer et al. [34] presented a novel RL approach to teaching another RL agent, referred to as the student, in the context of "teaching on a budget", where the teacher agent can only provide a limited number of suggestions. The state of the teaching task in this study includes information about the student's state and the current teaching session, and the teacher's actions represent the decision of whether to give advice or not. The results of the simulations show that the RL teacher was effective at adapting to when and how to provide advice based on the state of the student and the training session. Furthermore, the paper proposes an improved method for the student to exploit received advice through a max-update strategy, though it requires additional computation and modification of the student's architecture. The authors also identified potential venues for future research, including exploring the approach in different domains, evaluating its performance in near-optimal settings, and incorporating different types of feedback. Additionally, they suggest adding a classifier to predict the student's next action to simplify communication between the student and teacher agent.

Li et al. [35] proposed an adaptive learning system using a continuous latent trait model and a deep Q-learning algorithm. The system aims to improve learners' abilities to reach target levels. To address insufficient state transition data, a transition model estimator based on neural networks was developed to enhance the deep Q-learning algorithm's performance. Two simulation studies demonstrated the effectiveness of the proposed methodology in finding optimal learning policies without teacher assistance. Sánchez et al. [36] presented a practical and effective approach for designing distance education resources for children aged from three to five years. The resources are created by teachers, leveraging their understanding of student needs without requiring advanced computer skills. The paper introduced a curriculum project consisting of computer game-based resources for this age group, which received positive recognition from educational authorities and stakeholders in Valencia, Spain. Zhang et al. [37] introduced RLTLBO, an improved teaching-learning-based optimization (TLBO) algorithm that incorporates reinforcement learning (RL) techniques. The algorithm features a new learning mode that considers the influence of the teacher and utilizes a switching mechanism between different learning modes in the learner phase using Q-learning.

3.2.2. RL Techniques to Provide Hints and Quizzing

Another significant application of RL for education is training policies to provide hints within a task as feedback, particularly for complex areas such as high school algebra and visual programming to enhance student learning gains and engagement [38–41]. Previous research work [42] used the MDP framework to automatically generate hints for logic-proof tutoring by using students' historical data. Efremov et al. [43] used RL formalism without using students' historical data to train a hints policy for visual programming. He-Yueya et al. [44] studied how to use an RL-based policy to quiz students and understand their level of knowledge. These research studies are still in their preliminary stages and show the potential of RL-based policies to solve problems beyond the customization of student curricula. RL techniques are expected to yield a major contribution to providing students with responses for complex, open-ended assignments. An important research direction here is to train RL policies to balance different learning objectives while giving hints to students to complete the recent task.

3.2.3. RL Techniques for Adaptive Experimentation in Educational Platforms

Recently, RL methods have been increasingly used to assess education on online platforms. More specifically, a specific category of RL techniques named multi-armed bandits (MAB) is utilized in adaptive experimentation, where each student is assigned to a specific technology version and the algorithm observes the learning outcome of the student. Each incoming student will be assigned to the technology version that has been more effective for a former student [45–47]. Student personalization is not enabled in standard MAB algorithms. However, contextual MAB methods can include features for the students and tailor the assignments to enhance the student's learning gain. Whitehill et al. [17] studied the effect of the features of contextual MAB algorithms and discussed the effect of personalization on learning gains. Rafferty et al. [48] used a case study of adaptive experiments to send reminders about homework by email to students. In general, RL has potential in adaptive experimentation, and it is expected to be seen in real-world educational platforms. Fairness and equity must be given top priority in order to guarantee that all students have access to high-quality, individually tailored learning experiences that can help them realize their maximum potential.

3.2.4. RL Techniques for Instructional Sequencing in Education

Ronald Howard is a pioneer in proposing and using RL to adaptively order different educational activities to support the student education process, which is currently known as instructional sequencing [49]. It is well known that the order of instruction can have an impact on how effectively students learn, which attracted several researchers to address problems in this area of research [50]. For example, the authors of [51] utilized RL to enable the mathematical mechanism for formally optimizing the instructions sequence. Similarly, Atkinson [52] suggests that a theory of instruction consists of four steps: (1) modeling the process of learning, (2) specifying the permissible actions, (3) specifying the objectives, and (4) a measurement scale that allows values to be assigned to each action and the payoffs to achieving the objectives. These steps of the theory of instruction can be mapped to the MDP.

In [52], Atkinson explained that the transition function maps onto the learning process model in the context of instruction, where the students' states are the MDP states. Instructional activities are actions that can vary according to the cognitive states of the students. Such activities can be flashcards, problem steps, worked examples, problems, and game levels in the educational context. Finally, each instructional action can be assigned a specific cost to be used in the reward function.

An example of using an MDP in instructional sequencing is using two states for each knowledge component (KC). The two states for each KC are "correct" or "incorrect". The state of each KC is determined by the student's results for the questions, with a correct answer resulting in the "correct" state and an incorrect answer resulting in the "incorrect" state. The full state is represented by a binary vector of each individual KC state. The available actions for the student are a set of items to practice, each linked to a specific KC. The probability of transitioning from one state to another for each KC is outlined in a 2×2 transition matrix for each item. The objective is to help the student achieve the correct state for as many knowledge components (KCs) as feasible, as reflected in the reward function, which rewards a transition from "incorrect" to "correct" with 1, penalizes a transition from "correct" to "incorrect" with -1, and leaves a transition that remains unchanged with 0. The ideal instructional policy is simple: always present the item with the highest chance of transitioning the KC from the "incorrect" state to the "correct" state. The transition probabilities for each KC are the only components in this basic MDP model and can be determined by dividing the number of times that the students successfully transitioned from "incorrect" to "correct" by the number of instances where the students were given an item in the "incorrect" state. This calculation is referred to as the maximum likelihood transition probability. The illustrated MDP model assumes that the only objective is for students to answer the questions accurately. However, in reality, a student's ability to answer questions correctly does not always indicate a deep understanding, for instance, by chance or through incorrect reasoning. To consider the uncertainty of a student's true cognitive state, it would be necessary to use a partially observable Markov decision process (POMDP) [53].

In the domain of educational modeling and decision making, the POMDP framework stands as a powerful tool, utilizing a belief state space for factors that are not observed, including students' proficiency and knowledge. The agent in a POMDP is unaware of the state, but an observation function (O) links the states to the probability distributions of observations. In this example, whether a student provides a correct or incorrect response to a question is observed at each time step, with the likelihood of a correct or erroneous response based on the student's state for the knowledge component (KC) currently being taught. The two states for each KC are referred to as the learned state and the unlearned state, representing the student's grasp of the KC. Ignoring the reward function, this POMDP is equivalent to the Bayesian knowledge tracing (BKT) model, which has been utilized in intelligent tutoring systems for cognitive mastery learning [54,55].

BKT is not typically considered within the RL framework due to the absence of a specified reward function, although the use of BKT for mastery learning follows an implicit reward function. One possible reward function for cognitive mastery learning could be to receive a reward of 1 each time the estimated probability of the student learning a particular KC exceeds 95% and a reward of 0 otherwise. Under this function, items will be provided for a specific KC until confidence in the student's learning exceeds 95%, at which point the model will move on to the next KC. It should be noted that the optimal policy under this reward function, which represents cognitive mastery learning, may differ significantly from the optimal policy under alternative reward functions, such as receiving a reward of one for each KC that is actually in the learned state (which cannot be directly observed).

To effectively estimate the parameters of a POMDP, various algorithms come into play, including expectation maximization by Welch [56], spectral learning techniques as described by Hsu et al. [57] and Falakmasir et al. [58], or a straightforward grid search through the entire parameter space as proposed by Baker et al. [59].

In the domain of instructional sequencing within the literature of intelligent tutoring systems, two distinct approaches emerge: task loop adaptivity (outer loop) and step loop adaptivity (inner loop), concepts clarified by Vanlehn [38,60,61]. Task loop adaptivity involves the selection of distinct instructional activities or tasks by the RL agent, while step loop adaptivity requires the RL agent to decide on the specifics of each step in a fixed instructional task. An example of step loop adaptivity would be to decide if the solution to the next step should be shown or if the student should solve it themself, as noted by Chi et al. [62].

Unlocking the potential of adaptive learning in the realm of online education holds the promise of not only enhancing student outcomes but also lightening the workload for learners, instructors, and course designers. A pioneering study led by Bassen et al. [63] used the first RL model to dynamically schedule educational activities for a large online course using active learning. The model optimizes the sequence of course activities to maximize student outcomes while minimizing the number of assigned tasks. An experiment with over 1000 participants was conducted to assess the impact of this scheduling policy on learning gains, dropout rates, and student feedback. The results show that the RL model outperformed a traditional linear assignment condition in terms of learning gains and yielded similar results to a self-directed approach but with fewer activities and lower dropout rates.

3.2.5. RL Techniques for Modeling Students

Another approach involves employing RL to simulate the behavior of the student as opposed to the teacher. In this approach, the RL agent is the student, and the teacher represents the environment. This type of modeling is useful in an open-ended learning field where tasks are sequential, open-ended, and conceptual. Hence, this RL model can be used to diagnose students' mistakes and build an effective feedback environment [47,64]. Also, a student was used as a model to evaluate teaching methods. Similarly, the research work in [65] used students as RL agents to study the foundations of teaching for the purpose of sequential decision making. The curriculum design problem has been studied in various research topics [66–68] where the student is used as a learning agent. Rakhsha [69] investigated the issue of policy teaching when the student is used as an RL agent. The RL framework modeling a student as an agent is a critical research direction that can be used in future work. A major research question is how to integrate human-centered features of learning into RL agents so that the RL agents represent students properly.

3.2.6. RL Techniques for Generating Educational Content

RL approaches can be used to produce educational content such as videos, quizzes, and exercises. This type of content is known as procedural content generation (PCG). Research works studied the applicability of RL for PCG to generate various levels of racing games [70] and Sokoban puzzles [71,72]. The Monte Carlo tree search (MCTS) technique is used to generate new tasks in the visual programming field. These generated tasks can be used in various areas such as assigning new practice tasks, including quizzes or homework, to evaluate students' knowledge. If a student fails, then the student could be assigned a new task to aid in answering the given task. Minoofam et al. [73] developed RALF, an adaptive reinforcement learning framework based on Cellular Learning Automata (CLA), to automatically generate content for students with dyslexia. His work addresses the application of machine learning techniques in adaptive content generation for students with dyslexia. RL in the generation of educational content is an essential research field open for further investigation.

3.2.7. RL Techniques for Personalized Education through E-Learning

Information and communication technology (ICT) is becoming more significant in education as a result of global education reform. We expand the idea of personalization from e-commerce to education, referred to as personalized education (PE), to fully utilize the enormous amount of relevant and accessible hypermedia that can potentially be accessed on the Internet. PE includes identifying and comprehending the needs and competencies of each individual student before adopting and implementing the most effective teaching methods and multimedia content to close the knowledge gap. An agent-based framework for the construction of a personalized education system (PES) has been presented in order to deploy the right personalization strategies and fulfill the learning needs of a wide range of learners in various methods [74]. A PES is made up of diverse functional components that heavily rely on customization technology tailored to assisting with different learning activities. This incorporates intelligent user profiling, searching, and clustering of multimedia information as well as a constantly adapting intelligent user interface to each user's actions and interactions with the system. During childhood, individuals form a set of consistent interests, personalities, and skills as a result of both positive and negative cultural and educational experiences [75]. This reinforces the idea that PE is a high-impact area for the newest e-learning support systems. In [76], a formal method based on the hidden Markov model (HMM) is presented to learn and characterize user behavior from their interactions with a web-based learning system. The system creates and recognizes user profiles using the resulting HMM trained from several scenarios of user interactions in order to predict and anticipate user demands and dynamically change the user interface to meet the user's competence and learning needs.

Wu et al. [77] explored the use of reinforcement learning (RL) to develop data-driven policies for tutorial planning in adaptive training systems. The aim was to provide personalized scaffolding based on the Interactive, Constructive, Active, Passive framework for cognitive engagement. Tang et al. [37] discussed the concept of personalized learning, where instruction is tailored to individual learners' needs and optimized for their pace of learning. Recent advancements in information technology and data science have made personalized learning feasible through data-driven recommendation systems.

3.2.8. RL Techniques for Personalizing a Curriculum

One of the most extensively researched RL applications in education is to design an educational approach that can train an instructional policy to provide personalized learning materials to students. In such a situation, an RL agent is trained to generate an instructional policy in an intelligent tutoring system, with the student forming an integral part of the environment [9,78]. The instructional policy is responsible for keeping track of the student's response history and finding ways to optimize his or her long-term learning.

The intricacy of the prerequisite dependencies and the curriculum's organizational structure are essential elements that affect student progress and, in turn, graduation rates. However, we are not aware of any closed-form techniques for calculating the correlation between a curriculum's complexity and the percentage of students who successfully complete it. In [79], a novel approach is presented for quantifying this relationship using an MDP. The MDP is an appropriate method for addressing such a problem since student growth is non-deterministic and because their states change throughout each semester.

In [80], real university data, various data mining techniques, machine learning approaches, and graph theory were used to study the structure of courses. These strategies and procedures offer a quantitative instrument for measuring the complexity of a curriculum framework. The findings of this study demonstrate an inverse relationship between a curriculum's complexity and the percentage of students that successfully completed it.

In the literature, there are numerous definitions of student achievement. Most studies view graduation as the ultimate indicator of student achievement, although these factors might range from grades and perseverance to self-improvement [81]. The framework shown in Figure 5 is possibly the most typical one employed by colleges in the US when examining the elements influencing student achievement. Pre-institutional encounters, institutional circumstances, and student actions can be used to segment these characteristics into three broad categories [82,83]. Student growth will be facilitated by providing institutional settings that matter, but there might also be structural issues with the curricula that prevent advancement without reference to any success initiatives. The organization of the courses in a curriculum, such as the number of courses and prerequisites, determines the structural complexity of the curriculum. In contrast, a curriculum's instructional complexity is defined by the course's inherent difficulty, the caliber of the teacher, academic assistance, etc. These two factors define how challenging a curriculum could be. Graph theory and complex network analysis are employed to give a mathematical framework to identify essential courses. This may assist the university in deciding when to offer particular programs, who should teach them, and what is actually required for a degree in a particular discipline. The proposed method gives a strong framework to ensure that students may easily go through courses with the aim of increasing a university's graduation rates. A course is considered critical if the student cannot or does not take it at the appropriate time, thus failing to complete the course by the scheduled time. Therefore, it would be imperative to transfer these courses to the earliest terms while still adhering to the requirement relationship restrictions, maximum and minimum academic loads per term, maximum and minimum credit hours per term, etc. Pivotal courses have a significant influence on university student advancement and, eventually, graduation rates. Therefore, finding such courses should be a top priority for university decision makers. The blocking factor and delay factor are two key characteristics that relate to the cruciality of a course within the network [80]. The blocking factor of a course is the total number of classes that students are restricted from enrolling in until they successfully complete it. The delay factor of a course is the total number of prerequisite courses on any pathway that contains that course. In [79], an MDP was used for modeling curriculum analytics. The research provides a mathematical model that supports instantaneous what-if evaluations connected to curricular modifications rather than waiting years to assess the influence of curriculum changes on student progress.

Characterizing the Complexity of Curricular Patterns in Engineering Programs

Through the sophomore year, engineering programs often follow a similar pattern for teaching undergraduate students. Essentially, students who fail an earlier course in a curricular pattern or are unable to begin the pattern on time (for instance, because of problems with math placement) will frequently have to postpone graduation. Reforms to engineering curricula have been introduced in a number of schools with the goal of enhancing the percentage of students who graduate on time. In this study, the curricular analytics tools are used to evaluate the degree to which specific refinement will increase the rates of graduation. One of the most challenging tasks facing academic directors is altering the culture of an educational system or curriculum. This is because organizational culture is created over a long period of time by a complex web of objectives, procedures, customs, values, roles, and attitudes. Higher education has cultural components that have been developed over a thousand years of history.

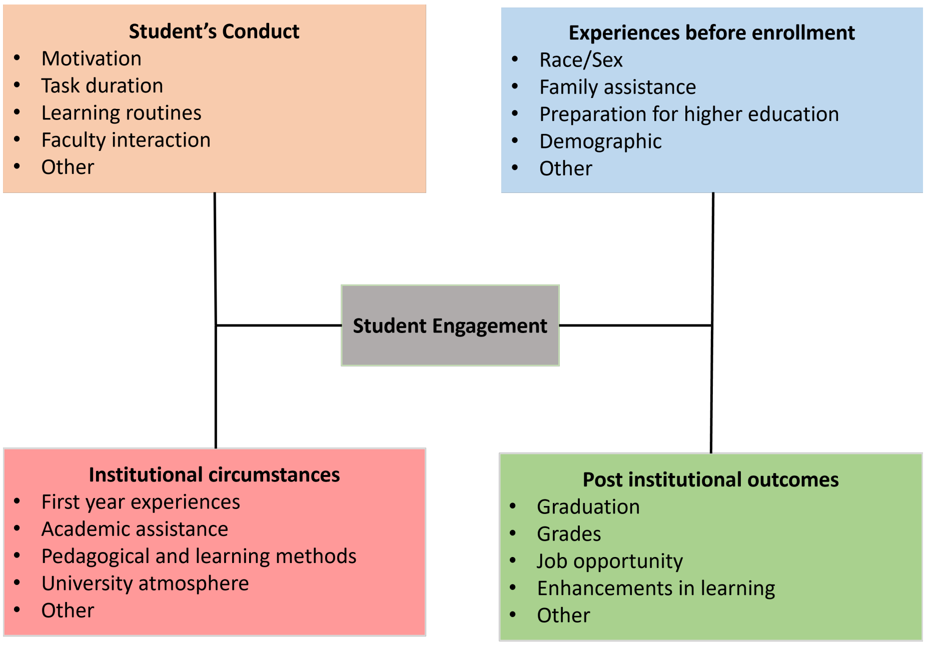

Figure 5. Student success framework.

In [84], the authors viewed the university as a system composed of a number of smaller units that work together to achieve the system's objectives. Each of these components has a role in the success of reform initiatives either directly or indirectly, given the fact that the characteristics of the university system vary from one university to another. The model that can be employed to forecast the projected benefits that can be achieved by adopting specific changes at specific universities would ideally be built before large-scale improvement initiatives are launched. This necessitates formalizing the university system model. A difficult aspect of a university's system formulation is determining the appropriate metrics and measurements that can be used to quantify the various subcomponents of the educational system. This is imperative if the system needs to be analyzed to make predictions, quantify the impact of interventions, and focus efforts where they are most likely to succeed.

In the following, formal approaches are discussed to define the significant patterns seen in engineering curricula and estimate their complexity. A demonstration of how the complexity measure can be used to analyze curricular patterns was carried out in [84]. It was noticed that there is a connection between a student's ability to navigate a given curricular pattern and the complexity of that pattern.

The design pattern shown in Figure 6 is one that many aspiring engineering students are asked to follow for the traditional Circuit I course. Precalculus becomes the most important course in the pattern due to its placement in the pattern. Each course in a curricular pattern contributes learning outcomes that are necessary to go through subsequent courses in the pattern, while some courses offer extra learning outcomes that are not necessary to complete the pattern (though they might be required in other parts of the curriculum). Hence, as long as all of the relevant learning outcomes needed to attempt the Circuits I subject occur in prerequisite courses, we can modify the order in which these objectives are attained.

This design pattern, represented in Figure 7, entails the replacement of Precalculus in the preceding design pattern with the introduction of a new course called Engineering 101. This course is based on the one developed at Wright State University, and the design pattern itself is based on the curricular changes made there. Notice that the only prerequisite for the Circuits I course is now Engineering 101. The learning outcomes remain unchanged. This design pattern tries to achieve the same learning outcomes as in Figure 6 with a markedly lower total pattern complexity for the same starting math preparation. This considerably raises a student's chances of finishing this pattern, based on a few quite basic assumptions. The pattern in Figure 6 requires five terms to complete, whereas this pattern may be finished in just four terms.

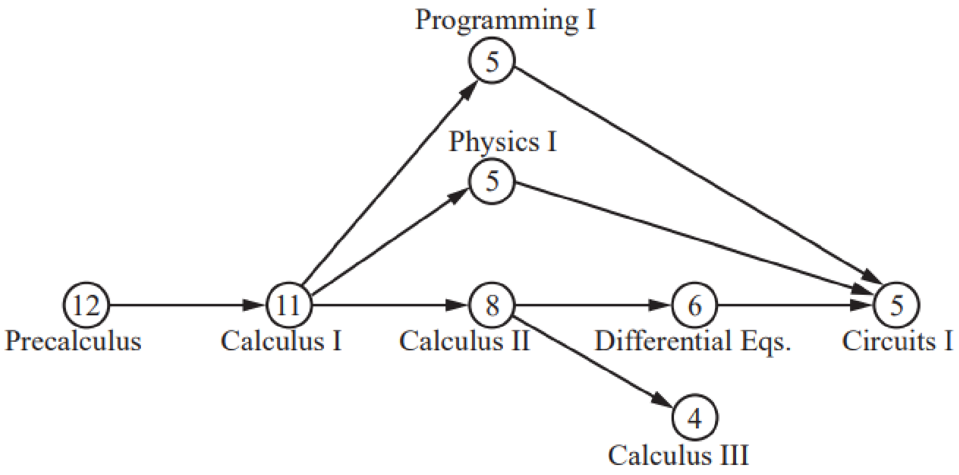

Figure 6. The traditional curricular pattern associated with Circuits I, with course cruciality shown inside the node [84].

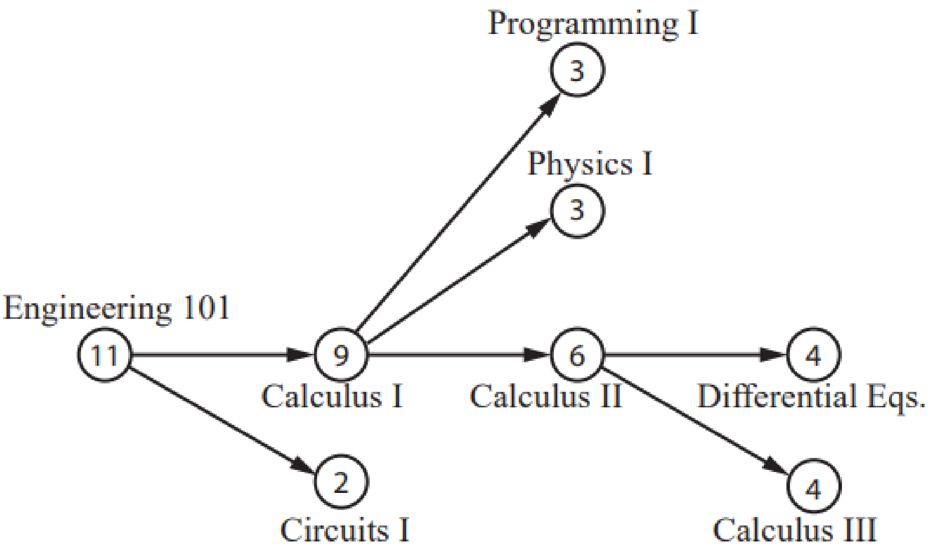

Figure 7. Alternative Circuits I design pattern [84].

Network Analysis of University Courses

Pivotal courses have a significant impact on how well university students perform and, eventually, on graduation rates. Therefore, finding such courses should be a top priority for university academic administrators. In [80], a new approach for identifying such courses and measuring their momentousness based on complex network analysis and graph theory is proposed. The proposed framework's applications are expanded to examine the complexity of an institution's curricula, which prompts thought about the development of ideal curricula. The best curricula are further used to analyze student progress, together with the achieved letter grades of the courses. This work is significant because it offers a solid framework to guarantee that students can easily proceed through curricula with the aim of raising a university's graduation rate [85]. Table 1 summarizes the RL applications and RL techniques discussed in the references incorporated in this literature review along with our study.

Table 1. RL applications and techniques in education field.

	RL Applications								RL Technique			
Ref.	**PC**	**TSF**	**H&Q**	**AE**	**IS**	**MS**	**GEC**	**PE**	**MDP**	**POMDP**	**DRL**	**MC**
[78]	✓				✓							
[9]	✓											
[11–14,30]											✓	
[79]	✓								✓			
[33,34]		✓										
[38]			✓		✓							
[39–41,43,44]			✓									
[42]			✓						✓			
[45,46,48]				✓								
[47]				✓		✓			✓			
[17]				✓						✓		
[18,49,86,87]					✓				✓			
[50–52,60–63,69,88]					✓							
[19]									✓			
[53–55,59]					✓					✓		
[56,57]										✓		
[64–68]						✓						
[70–72]							✓					
[74–76]								✓				
[31,32]												✓
[20–24]									✓			
[25–29]										✓		

The following abbreviations are used in this table: PC = personalized curriculum; TSF = teacher–student framework; H&Q = hints and quizzing; AE = adaptive experimentation in educational platforms; IS = instructional sequencing in education; MS = model students; GEC = generating educational content; PE = personalized education through e-learning; MDP = Markov decision process; POMDP = partially observable Markov decision process; DRL = deep RL framework; MC = Markov chain. Note: ✓ = does include.

4. Considerations in the Design of Reinforcement Learning

In reinforcement learning, it is vital to comprehend the various configurations and the design choices that must be made by researchers when utilizing RL. RL algorithms can be grouped into two categories: model-based and model-free. The model-based approach entails acquiring the transition and reward functions before utilizing MDP planning to generate a policy. Conversely, model-free methods directly learn the policy from data without the need to first learn the model. RL can be utilized in two ways: online and offline. In online RL, the agent adapts its policy as it interacts with the environment. On the other hand, in offline RL, a policy is learned from previously collected data and then applied to an actual environment. In online RL, the agent faces the challenge of balancing between using the current best policy to attain high rewards and trying out new, uncertain actions with the aim of discovering a better policy. This is referred to as the exploration-exploitation dilemma. Exploration involves attempting new actions to gather information from less familiar regions of the state and action space, while exploitation entails utilizing the best policy identified thus far. Although this trade-off is a crucial aspect in RL, it is rarely addressed in instructional sequencing studies [89–91]. Since the cognitive state of students is typically not visible, a POMDP is often utilized instead of a fully observable MDP. Solving POMDPs through planning or reinforcement learning is typically challenging, leading researchers to rely on approximate methods such as myopic planning [92]. However, some learning models, like the BKT model, are so limited that they can be used to find the optimal policy in a POMDP.

5. Challenges of Artificial Intelligence in Education and Future Research Directions

As discussed in the previous sections, AIEd has the potential to significantly enhance learning, teaching, evaluation, and academic administration by providing students with more individualized and adaptive learning, enhancing teachers' comprehension of students' learning processes and offering anywhere, anytime machine-supported inquiries and immediate feedback [93]. The success of AIEd is substantial; its advancement has also drawn attention to the opportunity for substantial and significant changes in education. These challenges require a lot of effort to be met. To develop proven AIEd tools and procedures, we must take advantage of and contribute to cognitive and AI research. The difficulties include combining the development of intelligent systems with empirical research, both in the laboratory and in the real world, to gather proof of the educational advantages of these tools, approaches, and systems. Some challenges and future research directions of AIEd are listed below:

- Insufficient pertinent learning resources for personalized or adaptive learning: Instructors have complained that the suggested pedagogies and learning materials of the personalized or adaptive learning platform are overly uniform. Learning objects are any standardized and reusable digital instructional resources that can be easily reused and customized to meet a learning purpose in a number of scenarios, according to the recommendations made by AI agents [94]. More research is required to better understand how learning objects should be used in customized and adaptive learning, as well as how they might be created.
- Lack of perspectives on education in AIEd research: The majority of AIEd researchers come from a strong engineering background; they frequently concentrate on technology design and development and approach AIEd research from an engineering perspective. The opinions of educational researchers and instructors are not adequately represented by this strategy. Future studies should look for innovative research techniques for diverse disciplines of AIEd that can directly involve educators, researchers, and students, as AI is an interdisciplinary field [95].
- Data selection for AI predictive models: The organized student information now utilized in linear regressions, a type of classical predictive model, is not necessarily suitable for nascent AI technology. An extensive collection of unstructured and structured student information is necessary for an efficient AI predictive model, which poses significant privacy concerns. The efficiency of AI technology must be balanced with ethical limitations because AIEd frequently targets young learners. Additional investigation is required on the types of information that should be used in AI models while taking ethical considerations very seriously [96].
- Socio-emotional factors are understudied in AIEd research: The majority of AIEd studies have focused on cognitive results and adaptive learning, whereas very few have looked at socio-emotional effects [97]. AIEd has been linked to risks and unfavorable outcomes, and both teachers and students are conscious of the moral issues involved [98–102]. The ethical implications of applying AI to the fields of social science, engineering, and law have not yet been thoroughly examined. Therefore, more investigation is required on the ethical problems surrounding AIEd.
- Teachers lack sufficient expertise in AI technologies: The majority of teachers have been instructing in a black box since they are unable to comprehend how AI technologies operate (for example, the guiding principles or algorithms for resource recommendations). Therefore, they cannot fully utilize the technologies for learning, teaching, and assessment, as well as respond to students' questions regarding AIEd (such as why particular learning resources were chosen by the AI platforms). Therefore, future studies should take into account the requirement for instructors to understand AI and its use in education [93].
- Potential ethical and social issues: In [103], four themes—privacy, replacing people, consequences for children, and responsibility—were used to analyze the ethical implications of deploying (humanoid) robots in the classroom. The nature of intelligence,

how to balance the interests of individuals and the general public, how to deal with moral conundrums, and how automation will affect the labor market are just a few of the fundamental concerns surrounding AI that cannot be fully addressed by technology. These issues necessitate interdisciplinary methods, which change the purpose and nature of educational programs.

For a complete grasp of the ethical and social implications of AI technology, it is crucial that educators and technologists offer guidance before fully integrating AI into education. Rather than fearing AI's world dominance, as some predict in the form of technological singularity [104], the primary concern should be people's readiness to simply trust the technology as a present, which results to practical issues such as mishandling, over-dependence, underutilization or ignorance, abuse, and use without consideration for the consequences [105]. The majority of the principles and guidelines that have emerged in the last few years from governments, policy entities, social agencies, and technology firms are strikingly devoid of specific recommendations for education, despite the fact that most of them acknowledge that the obviousness of education will be crucial to ensuring trustworthy and responsible AI [106].

In this age of artificial intelligence, there is a call for innovation and creativity. Alongside the acquisition of technological expertise, there is a growing need to instill skills such as imagination, collaboration, critical inquiry, and continuous learning into the fabric of education. This demonstrates how the old division between humanities, arts, social sciences, and STEM is inadequate for the demands of the new era of artificial intelligence. Future students must be transdisciplinary rather than just multidisciplinary in order to be competent in a range of intellectual frameworks outside of disciplinary views. AI is not just a STEM field [107,108]; rather, it is transdisciplinary in nature and necessitates a range of skills that are not taught in the curricula of today's schools. Redesigning studies is essential. The potential to fully attain inclusion and diversity across academic sectors will be provided by this. This will present an exceptional opportunity to actually achieve inclusion and diversity across academic sectors.

Combining data-driven and psychological approaches offers exciting potential for advancing instructional sequencing using RL. Psychological theories can guide the selection of models, action options, and strategies, as seen in past studies on learning tasks. It is advisable to focus on scenarios with limited yet meaningful choices and to apply educational principles, such as the expertise reversal effect, which suggests starting with worked examples and gradually reducing their use in favor of problem-solving tasks.

Integrating psychological theories and principles with data-driven approaches will be especially important as deep reinforcement learning continues to gain popularity in the field. Moreover, instructional sequencing can be combined with machine intelligence and human intelligence, such as incorporating student choice or teacher input.

RL has several potential applications in improving AIEd systems. One of the primary benefits of RL is its ability to enhance the personalization and adaptivity of AIEd systems by learning from student interactions and adapting to their individual needs. Additionally, RL can be used to recommend appropriate learning objects to each student based on their learning history and preferences. Furthermore, RL can be applied to model socio-emotional factors and personalize interventions based on a student's emotional state.

RL can also enhance the accuracy and efficiency of predictive models by learning from both structured and unstructured student data. However, it is essential to consider ethical implications, especially when dealing with young learners. Thus, RL can be utilized to balance the efficiency of AI technology with ethical limitations.

Moreover, RL algorithms can provide explanations for the recommendations made by AIEd systems, which can help teachers better understand and utilize these technologies. Additionally, RL can be employed to develop ethical guidelines and principles for AIEd systems, particularly in areas such as privacy, responsibility, and consequences for children. RL algorithms have the potential to address some of the challenges and research directions of AIEd, particularly in personalization, adaptivity, predictive modeling, and ethical considerations. However, it is crucial to balance the potential benefits of these technologies with the ethical and privacy concerns associated with the use of student data.

6. Discussion

In response to RQ1, the manuscript highlights that RL can be an effective framework for enhancing educational outcomes. Various studies have demonstrated positive results in utilizing RL in personalized learning, adaptive testing, and game-based learning within the education field. However, it also emphasizes the importance of acknowledging the limitations and challenges associated with implementing RL in educational settings. Regarding RQ2, the manuscript provides a comprehensive overview of the applications of RL in education. It discusses the use of RL in developing personalized curricula, generating hints and quizzing, generating educational content, adaptive testing, game-based learning, and curriculum design. This paper emphasizes the potential benefits of incorporating RL in these areas, including increased student engagement, improved learning outcomes, and reduced workload for teachers. As for RQ3, the manuscript illustrates the challenges and the future direction of RL in education, which involve the development of more advanced RL algorithms capable of handling complex learning scenarios, such as multi-agent systems. It also emphasizes the integration of RL with other AI techniques, such as natural language processing and computer vision. Additionally, this paper underscores the need for further research to fully comprehend the potential advantages and limitations of RL in education, establish best practices for its implementation, and explore the exciting potential of combining data-driven and psychological approaches in instructional sequencing using RL.

7. Conclusions

The objective of this literature review was to analyze the effectiveness and potential of RL applications and techniques in the field of education. The manuscript surveyed the various applications of RL in education to help come up with answers to the research questions. It investigated the best methods to use recent advancements in RL techniques to advance education technology, and researchers will be able to compare the usefulness and effectiveness of commonly employed RL applications and techniques in education, which will guide them in new directions. The review started by introducing the application of RL in education and explaining the different techniques involved, including the Markov decision process, partially observable Markov decision process, Markov chain, and deep RL framework.

Additionally, the literature review illustrated one of the specific topics within the application of RL in education: the restructuring of STEM program curricula based on the complexity and momentousness of courses and learning outcomes. The importance and challenges of this restructuring were discussed in detail, with a particular emphasis on the use of graph theory to calculate the complexity of the curriculum. The aim was to restructure the curriculum so that graduation rates improved without affecting the learning outcome. In conclusion, this literature review provides a comprehensive examination of the potential benefits and challenges of using RL algorithms in the field of education. By highlighting the various applications and techniques of RL in education, this review aims to answer the questions of how successful RL has been in facilitating education and what future directions the field may have. Finally, the main objective of this review was to provide both experienced and new researchers in the field with a thorough understanding of the use of RL in education, guiding future research and development in this area.

Author Contributions: Conceptualization, B.F.M. and A.W.; methodology, B.F.M. and A.W.; resources, B.F.M. and A.W.; writing—original draft preparation, B.F.M. and A.W.; writing—review and editing, M.H., A.S. and N.A.A.; supervision, M.H.; funding acquisition M.H. and N.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Big Data Analytics Center of United Arab Emirates University under grant code 12R135.

Data Availability Statement: We did not analyze or generate any datasets because our work proceeded within a theoretical approach. One can obtain the relevant materials from the references below.

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

References

1. Johri, A. Artificial intelligence and engineering education. *JEE* **2020**, *109*, 358–361. [CrossRef]
2. Besterfield-Sacre, M.; Shuman, L.J.; Wolfe, H.; Clark, R.M.; Yildirim, P. Development of a work sampling methodology for behavioral observations: Application to teamwork. *J. Eng. Educ.* **2007**, *96*, 347–357. [CrossRef]
3. Butz, B.P. The learning mechanism of the interactive multimedia intelligent tutoring system (IMITS). *J. Eng. Educ.* **2001**, *90*, 543–548. [CrossRef]
4. Fahd, K.; Venkatraman, S.; Miah, S.; Ahmed, K. Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Educ. Inf. Technol.* **2022**, *27*, 1–33. [CrossRef]
5. Qazdar, A.; Er-Raha, B.; Cherkaoui, C.; Mammass, D. A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco. *Educ. Inf. Technol.* **2019**, *24*, 3577–3589. [CrossRef]
6. Liu, X.; Ardakani, S. A machine learning enabled affective E-learning system model. *Educ. Inf. Technol.* **2022**, *27*, 9913–9934. [CrossRef]
7. Wiering, M.A.; Van Otterlo, M. Reinforcement learning. *Adapt. Learn. Optim.* **2012**, *12*, 729.
8. Puterman, M.L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*; John Wiley & Sons: Hoboken, NJ, USA, 2014.
9. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
10. Kaelbling, L.P.; Littman, M.L.; Moore, A.W. Reinforcement learning: A survey. *J. Artif. Intell. Res.* **1996**, *4*, 237–285. [CrossRef]
11. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv* **2013**, arXiv:1312.5602.
12. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef]
13. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef]
14. Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; et al. Mastering the game of go without human knowledge. *Nature* **2017**, *550*, 354–359. [CrossRef] [PubMed]
15. Mothanna, Y.; Hewahi, N. Review on Reinforcement Learning in CartPole Game. In Proceedings of the 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), Sakheer, Bahrain, 20–21 November 2022; pp. 344–349. [CrossRef]
16. Souchleris, K.; Sidiropoulos, G.K.; Papakostas, G.A. Reinforcement Learning in Game Industry—Review, Prospects and Challenges. *Appl. Sci.* **2023**, *13*, 2443. [CrossRef]
17. Whitehill, J.; Movellan, J. Approximately optimal teaching of approximately optimal learners. *IEEE Trans. Learn. Technol.* **2017**, *11*, 152–164. [CrossRef]
18. Sutton, R.S.; Barto, A.G. *Introduction to Reinforcement Learning*; MIT Press: Cambridge, MA, USA, 1998.
19. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*; Elsevier: Amsterdam, The Netherlands, 1994; pp. 157–163.
20. Iglesias, A.; Martínez, P.; Aler, R.; Fernández, F. Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning. *Appl. Intell.* **2009**, *31*, 89–106. [CrossRef]
21. Iglesias, A.; Martínez, P.; Aler, R.; Fernández, F. Reinforcement learning of pedagogical policies in adaptive and intelligent educational systems. *Knowl.-Based Syst.* **2009**, *22*, 266–270. [CrossRef]
22. Iglesias, A.; Martinez, P.; Fernández, F. An experience applying reinforcement learning in a web-based adaptive and intelligent educational system. *Inform. Educ.* **2003**, *2*, 223–240. [CrossRef]
23. Martin, K.N.; Arroyo, I. AgentX: Using reinforcement learning to improve the effectiveness of intelligent tutoring systems. In Proceedings of the International Conference on Intelligent Tutoring Systems, Maceió, Brazil, 30 August–3 September 2004; Springer: Berlin/Heidelberg, Germany, 2004; pp. 564–572.
24. Chi, M.; VanLehn, K.; Litman, D.; Jordan, P. Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies. *User Model. User-Adapt. Interact.* **2011**, *21*, 137–180. [CrossRef]
25. Jaakkola, T.; Singh, S.; Jordan, M. Reinforcement learning algorithm for partially observable Markov decision problems. *Adv. Neural Inf. Process. Syst.* **1994**, *7, 345–352.*
26. Koenig, S.; Simmons, R. Xavier: A robot navigation architecture based on partially observable markov decision process models. In *Artificial Intelligence Based Mobile Robotics: Case Studies of Successful Robot Systems*; MIT Press: Cambridge, MA, USA, 1998; pp. 91–122.
27. Mandel, T.; Liu, Y.E.; Levine, S.; Brunskill, E.; Popovic, Z. Offline policy evaluation across representations with applications to educational games. In Proceedings of the AAMAS, Paris, France, 5–9 May 2014; Volume 1077.

28. Rafferty, A.N.; Brunskill, E.; Griffiths, T.L.; Shafto, P. Faster teaching via pomdp planning. *Cogn. Sci.* **2016**, *40*, 1290–1332. [CrossRef]
29. Clement, B.; Oudeyer, P.Y.; Lopes, M. A Comparison of Automatic Teaching Strategies for Heterogeneous Student Populations. In Proceedings of the International Educational Data Mining Society, Raleigh, North Carolina, 29 June–2 July 2016.
30. Wang, P.; Rowe, J.P.; Min, W.; Mott, B.W.; Lester, J.C. Interactive Narrative Personalization with Deep Reinforcement Learning. In Proceedings of the IJCAI, Melbourne, Australia, 19–25 August 2017; pp. 3852–3858.
31. Luo, M. Application of AHP-DEA-FCE model in college English teaching quality evaluation. *Int. J. Appl. Math. Stat.* **2013**, *51*, 101–108.
32. Yuan, T. Algorithm of classroom teaching quality evaluation based on Markov chain. *Complexity* **2021**, *2021*, 1–12. [CrossRef]
33. Anand, D.; Gupta, V.; Paruchuri, P.; Ravindran, B. An enhanced advising model in teacher-student framework using state categorization. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 6653–6660.
34. Zimmer, M.; Viappiani, P.; Weng, P. Teacher-student framework: A reinforcement learning approach. In Proceedings of the AAMAS Workshop Autonomous Robots and Multirobot Systems, Paris, France, 5–9 May 2014.
35. Li, X.; Xu, H.; Zhang, J.; Chang, H.h. Deep reinforcement learning for adaptive learning systems. *J. Educ. Behav. Stat.* **2023**, *48*, 220–243. [CrossRef]
36. Tárraga-Sánchez, M.d.l.Á.; Ballesteros-García, M.d.M.; Migallón, H. Teacher-Developed Computer Games for Classroom and Online Reinforcement Learning for Early Childhood. *Educ. Sci.* **2023**, *13*, 108. [CrossRef]
37. Tang, X.; Chen, Y.; Li, X.; Liu, J.; Ying, Z. A reinforcement learning approach to personalized learning recommendation systems. *Br. J. Math. Stat. Psychol.* **2019**, *72*, 108–135. [CrossRef]
38. Aleven, V.; McLaughlin, E.A.; Glenn, R.A.; Koedinger, K.R. Instruction based on adaptive learning technologies. In *Handbook of Research on Learning and Instruction*; Routledge: New York, NY, USA, 2016; pp. 522–560.
39. Williams, J.J.; Kim, J.; Rafferty, A.; Maldonado, S.; Gajos, K.Z.; Lasecki, W.S.; Heffernan, N. Axis: Generating explanations at scale with learnersourcing and machine learning. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale, Edinburgh, UK, 25–26 April 2016; pp. 379–388.
40. Patikorn, T.; Heffernan, N.T. Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In Proceedings of the Seventh ACM Conference on Learning@ Scale, Virtual, USA, 12–14 August 2020; pp. 115–124.
41. Erickson, J.A.; Botelho, A.F.; McAteer, S.; Varatharaj, A.; Heffernan, N.T. The automated grading of student open responses in mathematics. In Proceedings of the Tenth International Conference on Learning Analytics & Knowledge, Frankfurt, Germany, 23–27 March 2020; pp. 615–624.
42. Barnes, T.; Stamper, J. Toward automatic hint generation for logic proof tutoring using historical student data. In Proceedings of the International Conference on Intelligent Tutoring Systems, Montreal, QC, Canada, 23–27 June 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 373–382.
43. Efremov, A.; Ghosh, A.; Singla, A. Zero-shot learning of hint policy via reinforcement learning and program synthesis. In Proceedings of the EDM, Virtual, 10–13 July 2020.
44. He-Yueya, J.; Singla, A. Quizzing Policy Using Reinforcement Learning for Inferring the Student Knowledge State. *Int. Educ. Data Min. Soc.* **2021**, 533–539.
45. Liu, Y.E.; Mandel, T.; Brunskill, E.; Popovic, Z. Trading Off Scientific Knowledge and User Learning with Multi-Armed Bandits. In Proceedings of the EDM, London, UK, 4–7 July 2014; pp. 161–168.
46. Williams, J.J.; Rafferty, A.N.; Tingley, D.; Ang, A.; Lasecki, W.S.; Kim, J. Enhancing online problems through instructor-centered tools for randomized experiments. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–12.
47. Rafferty, A.N.; Ying, H.; Williams, J.J. Bandit assignment for educational experiments: Benefits to students versus statistical power. In Proceedings of the International Conference on Artificial Intelligence in Education, London, UK, 27–30 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 286–290.
48. Rafferty, A.; Ying, H.; Williams, J. Statistical consequences of using multi-armed bandits to conduct adaptive educational experiments. *J. Educ. Data Min.* **2019**, *11*, 47–79.
49. Howard, R.A. *Dynamic Programming and Markov Processes*; MIT Press: Cambridge, MA, USA, 1960.
50. Ritter, F.E.; Nerb, J.; Lehtinen, E.; O'Shea, T.M. *In Order to Learn: How the Sequence of Topics Influences Learning*; Oxford University Press: Oxford, UK, 2007.
51. Atkinson, R.C. Ingredients for a theory of instruction. *Am. Psychol.* **1972**, *27*, 921. [CrossRef]
52. Atkinson, R.C. Optimizing the learning of a second-language vocabulary. *J. Exp. Psychol.* **1972**, *96*, 124. [CrossRef]
53. Sondik, E.J. *The Optimal Control of Partially Observable Markov Processes*; Stanford University: Stanford, CA, USA, 1971.
54. Corbett, A.T.; Anderson, J.R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.* **1994**, *4*, 253–278. [CrossRef]
55. Corbett, A. Cognitive mastery learning in the act programming tutor. In Proceedings of the Adaptive User Interfaces. AAAI SS-00-01. 2000. Available online: https://api.semanticscholar.org/CorpusID:16877673 (accessed on 14 February 2023).
56. Welch, L.R. Hidden Markov models and the Baum-Welch algorithm. *IEEE Inf. Theory Soc. Newsl.* **2003**, *53*, 10–13.

57. Hsu, D.; Kakade, S.M.; Zhang, T. A spectral algorithm for learning hidden Markov models. *J. Comput. Syst. Sci.* **2012**, *78*, 1460–1480. [CrossRef]
58. Falakmasir, M.H.; Pardos, Z.A.; Gordon, G.J.; Brusilovsky, P. A Spectral Learning Approach to Knowledge Tracing. In Proceedings of the EDM, Memphis, TN, USA, 6–9 July 2013; pp. 28–34.
59. Baker, R.S.d.; Corbett, A.T.; Gowda, S.M.; Wagner, A.Z.; MacLaren, B.A.; Kauffman, L.R.; Mitchell, A.P.; Giguere, S. Contextual slip and prediction of student performance after use of an intelligent tutor. In Proceedings of the User Modeling, Adaptation, and Personalization: 18th International Conference, UMAP 2010, Big Island, HI, USA, 20–24 June 2010; Proceedings 18; Springer: Berlin/Heidelberg, Germany, 2010; pp. 52–63.
60. VanLehn, K. The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* **2006**, *16*, 227–265.
61. VanLehn, K. Regulative loops, step loops and task loops. *Int. J. Artif. Intell. Educ.* **2016**, *26*, 107–112. [CrossRef]
62. Chi, M.; Jordan, P.W.; Vanlehn, K.; Litman, D.J. To elicit or to tell: Does it matter? In Proceedings of the Aied, Brighton, UK, 6–10 July 2009; pp. 197–204.
63. Bassen, J.; Balaji, B.; Schaarschmidt, M.; Thille, C.; Painter, J.; Zimmaro, D.; Games, A.; Fast, E.; Mitchell, J.C. Reinforcement learning for the adaptive scheduling of educational activities. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–12.
64. Yang, X.; Zhou, G.; Taub, M.; Azevedo, R.; Chi, M. Student Subtyping via EM-Inverse Reinforcement Learning. *Int. Educ. Data Min. Soc.* **2020**, 269–279.
65. Zhu, X.; Singla, A.; Zilles, S.; Rafferty, A.N. An overview of machine teaching. *arXiv* **2018**, arXiv:1801.05927.
66. Haug, L.; Tschiatschek, S.; Singla, A. Teaching inverse reinforcement learners via features and demonstrations. *Adv. Neural Inf. Process. Syst.* **2018**, *31, 8464–8473.*
67. Tschiatschek, S.; Ghosh, A.; Haug, L.; Devidze, R.; Singla, A. Learner-aware teaching: Inverse reinforcement learning with preferences and constraints. *Adv. Neural Inf. Process. Syst.* **2019**, *32*. Available online: https://proceedings.neurips.cc/paper_files/paper/2019/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html (accessed on 14 February 2023).
68. Kamalaruban, P.; Devidze, R.; Cevher, V.; Singla, A. Interactive teaching algorithms for inverse reinforcement learning. *arXiv* **2019**, arXiv:1905.11867.
69. Rakhsha, A.; Radanovic, G.; Devidze, R.; Zhu, X.; Singla, A. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 7974–7984.
70. Gisslén, L.; Eakins, A.; Gordillo, C.; Bergdahl, J.; Tollmar, K. Adversarial reinforcement learning for procedural content generation. In Proceedings of the 2021 IEEE Conference on Games (CoG), Copenhagen, Denmark, 17–20 August 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–8.
71. Khalifa, A.; Bontrager, P.; Earle, S.; Togelius, J. Pcgrl: Procedural content generation via reinforcement learning. In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, Virtual, 19–23 October 2020; Volume 16, pp. 95–101.
72. Kartal, B.; Sohre, N.; Guy, S.J. Data driven Sokoban puzzle generation with Monte Carlo tree search. In Proceedings of the Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference, Burlingame, CA, USA, 8–12 October 2016.
73. Minoofam, S.A.H.; Bastanfard, A.; Keyvanpour, M.R. RALF: An adaptive reinforcement learning framework for teaching dyslexic students. *Multimed. Tools Appl.* **2022**, *81*, 6389–6412. [CrossRef]
74. Fok, A.W.P.; Ip, H.H. Personalized Education (PE) œ Technology Integration for Individual Learning. 2004. Available online https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0e4d60d16aec5ca0202f59957161c9a91a50d56a (accessed on 2 February 2023).
75. Ackerman, P.L. *Traits and Knowledge as Determinants of Learning and Individual Differences: Putting It All Together*; American Psychological Association: Washington, DC, USA, 1999.
76. Fok, A.W.; Wong, H.S.; Chen, Y. Hidden Markov model based characterization of content access patterns in an e-learning environment. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6–9 July 2005; IEEE: Piscataway, NJ, USA, 2005; pp. 201–204.
77. Wu, D.; Wang, S.; Liu, Q.; Abualigah, L.; Jia, H. An improved teaching-learning-based optimization algorithm with reinforcement learning strategy for solving optimization problems. *Comput. Intell. Neurosci.* **2022**, *2022.* [CrossRef] [PubMed]
78. Durik, A.M.; Hulleman, C.S.; Harackiewicz, J.M. One size fits some: Instructional enhancements to promote interest. In *Interest in Mathematics and Science Learning*; American Educational Research Association location: Washington, DC, USA, 2015; pp. 49–62.
79. Slim, A.; Al Yusuf, H.; Abbas, N.; Abdallah, C.T.; Heileman, G.L.; Slim, A. A Markov Decision Processes Modeling for Curricular Analytics. In Proceedings of the 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), Virtually Online, 13–15 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 415–421.
80. Slim, A. Curricular Analytics in Higher Education. Ph.D. Thesis, The University of New Mexico, Albuquerque, NM, USA, 2016.
81. Venezia, A.; Callan, P.M.; Finney, J.E.; Kirst, M.W.; Usdan, M.D. The Governance Divide : A Report on a Four-State Study on Improving College Readiness and Success. National Center Report# 05-3. National Center for Public Policy and Higher Education 2005. Available online https://eric.ed.gov/?id=ED508097 (accessed on 10 February 2023)
82. Whitt, E.J.; Schuh, J.H.; Kinzie, J.; Kuh, G.D. *Student Success in College: Creating Conditions That Matter*; Jossey-Bass: Hoboken, NJ, USA, 2013.

83. Tinto, V. *Leaving College: Rethinking the Causes and Cures of Student Attrition*; University of Chicago Press: Chicago, IL, USA, 2012.
84. Heileman, G.L.; Hickman, M.; Slim, A.; Abdallah, C.T. Characterizing the complexity of curricular patterns in engineering programs. In Proceedings of the 2017 ASEE Annual Conference & Exposition, Columbus, OH, USA, 25–28 June 2017.
85. Slim, A.; Kozlick, J.; Heileman, G.L.; Wigdahl, J.; Abdallah, C.T. Network analysis of university courses. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Republic of Korea, 7–11 April 2014; pp. 713–718.
86. Yan, H.; Yu, C. Repair of full-thickness cartilage defects with cells of different origin in a rabbit model. *Arthrosc. J. Arthrosc. Relat. Surg.* **2007**, *23*, 178–187. [CrossRef] [PubMed]
87. Ekinci, Y.; Ülengin, F.; Uray, N.; Ülengin, B. Analysis of customer lifetime value and marketing expenditure decisions through a Markovian-based model. *Eur. J. Oper. Res.* **2014**, *237*, 278–288. [CrossRef]
88. Bellman, R. A Markovian decision process. *J. Math. Mech.* **1957**, 6, 679–684. [CrossRef]
89. Lindsey, R.V.; Mozer, M.C.; Huggins, W.J.; Pashler, H. Optimizing instructional policies. *Adv. Neural Inf. Process. Syst.* **2013**, *26, 2778–2786.*
90. Clement, B.; Roy, D.; Oudeyer, P.Y.; Lopes, M. Multi-armed bandits for intelligent tutoring systems. *arXiv* **2013**, arXiv:1310.3174.
91. Segal, A.; Ben David, Y.; Williams, J.J.; Gal, K.; Shalom, Y. Combining difficulty ranking with multi-armed bandits to sequence educational content. In Proceedings of the Artificial Intelligence in Education: 19th International Conference, AIED 2018, London, UK, 27–30 June 2018; Proceedings, Part II 19; Springer: Berlin/Heidelberg, Germany, 2018; pp. 317–321.
92. Matheson, J.E. *Optimum Teaching Procedures Derived from Mathematical Learning Models*; Stanford University, Institute in Engineering-Economic Systems: Stanford, CA, USA, 1964.
93. Xia, Q.; Chiu, T.K.; Zhou, X.; Chai, C.S.; Cheng, M. Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Comput. Educ. Artif. Intell.* **2022**, 100118. . [CrossRef]
94. Cao, J.; Yang, T.; Lai, I.K.W.; Wu, J. Student acceptance of intelligent tutoring systems during COVID-19: The effect of political influence. *Int. J. Electr. Eng. Educ.* **2021**. . [CrossRef]
95. Holstein, K.; McLaren, B.M.; Aleven, V. Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. *Grantee Submiss.* **2019**. [CrossRef]
96. Sharma, K.; Papamitsiou, Z.; Giannakos, M. Building pipelines for educational data using AI and multimodal analytics: A "grey-box" approach. *Br. J. Educ. Technol.* **2019**, *50*, 3004–3031. [CrossRef]
97. Salas-Pilco, S.Z. The impact of AI and robotics on physical, social-emotional and intellectual learning outcomes: An integrated analytical framework. *Br. J. Educ. Technol.* **2020**, *51*, 1808–1825. [CrossRef]
98. Wood, E.A.; Ange, B.L.; Miller, D.D. Are we ready to integrate artificial intelligence literacy into medical school curriculum: Students and faculty survey. *J. Med. Educ. Curric. Dev.* **2021**, *8*, 23821205211024078. [CrossRef] [PubMed]
99. Kahn, K.; Winters, N. Constructionism and AI: A history and possible futures. *Br. J. Educ. Technol.* **2021**, *52*, 1130–1142. [CrossRef]
100. Banerjee, M.; Chiew, D.; Patel, K.T.; Johns, I.; Chappell, D.; Linton, N.; Cole, G.D.; Francis, D.P.; Szram, J.; Ross, J.; et al. The impact of artificial intelligence on clinical education: Perceptions of postgraduate trainee doctors in London (UK) and recommendations for trainers. *BMC Med. Educ.* **2021**, *21*, 1–10. [CrossRef] [PubMed]
101. Haseski, H.I. What Do Turkish Pre-Service Teachers Think About Artificial Intelligence? *Int. J. Comput. Sci. Educ. Sch.* **2019**, *3*, 3–23. [CrossRef]
102. Parapadakis, D. Can Artificial Intelligence Help Predict a Learner's Needs? Lessons from Predicting Student Satisfaction. *Lond. Rev. Educ.* **2020**, *18*, 178–195. [CrossRef]
103. Serholt, S.; Barendregt, W.; Vasalou, A.; Alves-Oliveira, P.; Jones, A.; Petisca, S.; Paiva, A. The case of classroom robots: Teachers' deliberations on the ethical tensions. *AI Soc.* **2017**, *32*, 613–631. [CrossRef]
104. Bostrom, N. The control problem. Excerpts from superintelligence: Paths, dangers, strategies. In *Science Fiction and Philosophy: From Time Travel to Superintelligence*; Wiley-Blackwell: Hoboken, NJ, USA, 2016; pp. 308–330.
105. Parasuraman, R.; Riley, V. Humans and automation: Use, misuse, disuse, abuse. *Hum. Factors* **1997**, *39*, 230–253. [CrossRef]
106. Dignum, V. The role and challenges of education for responsible AI. *Lond. Rev. Educ.* **2021**, *19*, 1–11. [CrossRef]
107. Dignum, V. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*; Springer: Berlin/Heidelberg, Germany, 2019.
108. Dignum, V. AI is multidisciplinary. *AI Matters* **2020**, *5*, 18–21. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.