Multicriteria Algorithms with Applications

Journal Homepage: sciencesforce.com/mawa

Multicriteria Algo. Appl. Vol. 4 (2024) 79–92

Paper Type: Review Article

Reinforcement Learning in Social Sciences: A Survey

Doaa El-Shahat [1,*], Nourhan Talal [1] and Mohamed Abouhawwash [2,3]

[1] Department of Computer Science, Faculty of Computer and Informatics, Zagazig University, Zagazig, 44519, Egypt. Emails: doaazidan@zu.edu.eg; N.Talal22@fci.zu.edu.eg.

[2] Department of Computational Mathematics, Science, and Engineering (CMSE), Michigan State University, East Lansing, United States; abouhaww@msu.edu.

[3] Department of Mathematics, Faculty of Science, Mansoura University, Mansoura 35516, Egypt.

Received: 07 Mar 2024 **Revised**: 04 Jul 2024 **Accepted**: 01 Aug 2024 **Published**: 05 Aug 2024

Abstract

Reinforcement Learning (RL) has become one of the most prominent topics in artificial intelligence research. It is widely used in various fields, such as recommendation systems, psychology, economics, and natural language dialogue systems. Finding the best path of action to maximize cumulative reward is the long-term strategy of RL. Undertaking research may yield suboptimal immediate results but optimal long-term consequences. Economists can address difficult behavioral problems with knowledge, especially those generated by deep learning algorithms. We provide the most recent advancements in RL methods in this study, along with their applications in gaming, finance, and economics. The survey's last section discusses RL's present problems and potential future developments. Such open problems as sample efficiency, safety, and interpretability are currently being sought after by researchers. Moreover, several ambitious prospective applications of RL in a wide variety of domains are discussed. This study gives a comprehensive review of the many methods and uses of RL in social science. This study's results will give researchers a standard against which to evaluate the utility and efficacy of frequently used RL. Guide future investigations across several domains.

Keywords: Reinforcement Learning; Model-based Reinforcement Learning; Model-free Reinforcement Learning; Markov Decision Processes; Artificial Intelligence; Social Sciences.

1 | Introduction

RL is a learning process where an artificial intelligence (AI) agent interacts with its environment through trial and error, acquiring the optimal behavioral strategy from rewards received in previous interactions. RL is the broad problem of learning behavior to optimize a long-term performance metric in a sequential setting. RL approaches may be used to solve goal-directed or optimization issues that can be converted to sequential decision-making problems. As a result, RL is closely related to optimal control and operations research, with strong links to optimization, statistics, game theory, causal inference, sequential experimentation, and other fields, and is useful to a wide range of challenges in science, engineering, and the arts [1].

RL allows computers to learn through real-world interaction. RL, to put it briefly, divides the real world into two parts: an environment and a representative. Through certain activities, the agent engages with the environment, and the environment provides feedback to the agent. In RL, the feedback is commonly referred

Corresponding Author: doaazidan@zu.edu.eg

https://doi.org/10.61356/j.mawa.2024.4353

to as the "reward." By attempting to obtain more favorable rewards from the surroundings, the agent can function "better." Through the use of RL algorithms, this learning process creates a feedback loop between the agent and the environment that directs the agent's progress[2].

RL [3, 4] is a machine-learning technique that focuses on how an intelligent agent interacts with its surroundings. It is useful for sequential decision-making since it learns the policy through trial and error search. As a result, it may offer viable ways to represent how a user and agent interact. Specifically, Deep Reinforcement Learning (DRL) [5], which combines deep learning techniques with classical RL, can learn from historical data with vast state and action spaces to solve large-scale issues. Its strong representation learning and function approximation capabilities may be used in a variety of contexts [6, 7], such as robots [8] and gaming [9].

A recent development in recommender system research is the use of RL to address recommendation difficulties [10-12]. In particular, RL allows the recommender agent to continuously suggest products to consumers to figure out the best recommendation strategies [13, 14]. Several experimental findings have shown that supervised learning approaches are inferior to RL-based recommendation systems.

Markov decision processes (MDP) are commonly used in RL to optimize policies [15]. The goal of the RL agent is to maximize the expected long-term return for each state. In MDP, estimating the value function of states and actions requires considering their transition probability. To estimate the value function of the states and actions, it is crucial to know the transition probability of the states in MDP. However, in many RL optimization situations, the model of the transition probability is not precisely measurable. As a result, model-free reinforcement [16, 17] learning techniques are widely used to find the RL agents' ideal policies. The historical trajectories produced by the agent's current policy are used to compute policy assessment and improvement in the absence of the transition probability model.

However, utilizing the learned model, model-based reinforcement [18, 19] learning techniques may replicate the state changes. As a result, how the surroundings and the agents interact can be prevented, leading to a higher sampling efficiency. However, the transition probability model is frequently computed using statistically erroneous historical data from a particular context. Many current practice applications overlook model flaws and train the agent's policy using the learned model as the real transition probability, which influences the taught policy. Next, the model-learned optimum policy is highly susceptible to changes in the transition probability and might potentially cause significant issues with real-world performance.

The agent and environment are essential components of RL. One domain for agent interaction is the environment. The goal of RL algorithms is to teach the agent how to interact with the environment in a way that will allow it to score highly on a predetermined criterion. In Pong, for example, the measure might be represented by scoring points. The agent receives a reward of one when the ball hits the other wall. On the other hand, the opponent receives a reward of one if the agent misses the ball and touches its wall [20].

In interactive RL, human input is employed either alone or in conjunction with external rewards. A few of the applications for RL are shown in Figure 1. There are several methods to integrate human input with RL, including evaluation [21], corrective [22], and guiding feedback [23].

Li et al. [24] Talk about different interpretations of human evaluative feedback in interactive RL. They distinguish three types of human evaluation input: learning from policy feedback, learning from category feedback, and interactive shaping. In interactive shaping, human feedback is interpreted as numerical incentives.

In contrast to other studies, we highlight the link and growth tendency by thoroughly reviewing various RL in the social sciences rather than concentrating on a single field. Moreover, we offer viable approaches to tackle the problems using RL and methodically classify the sophisticated RL techniques and their uses.

1.1 | Related Studies and Contribution of this Work

Many RL approaches have been masterfully used in many problems, as the literature now in publication attests. On RL, several writers have published survey and review works. Table 1 displays a synopsis of review papers that have been published so far on RL methodologies.

Since there are no penalties for bad behavior and data is almost limitless in simulation, the majority of developments have occurred in online RL. It was very difficult to apply these methods to the real world because many interesting systems are usually too complex to imitate [25]. Being able to learn a policy using pre-collected data without risk or expense to engage with the real world [26] is one of the attractions of learning offline enhancement [27], self-driving [28], health care [29, 30], dialogue systems [31], and others.

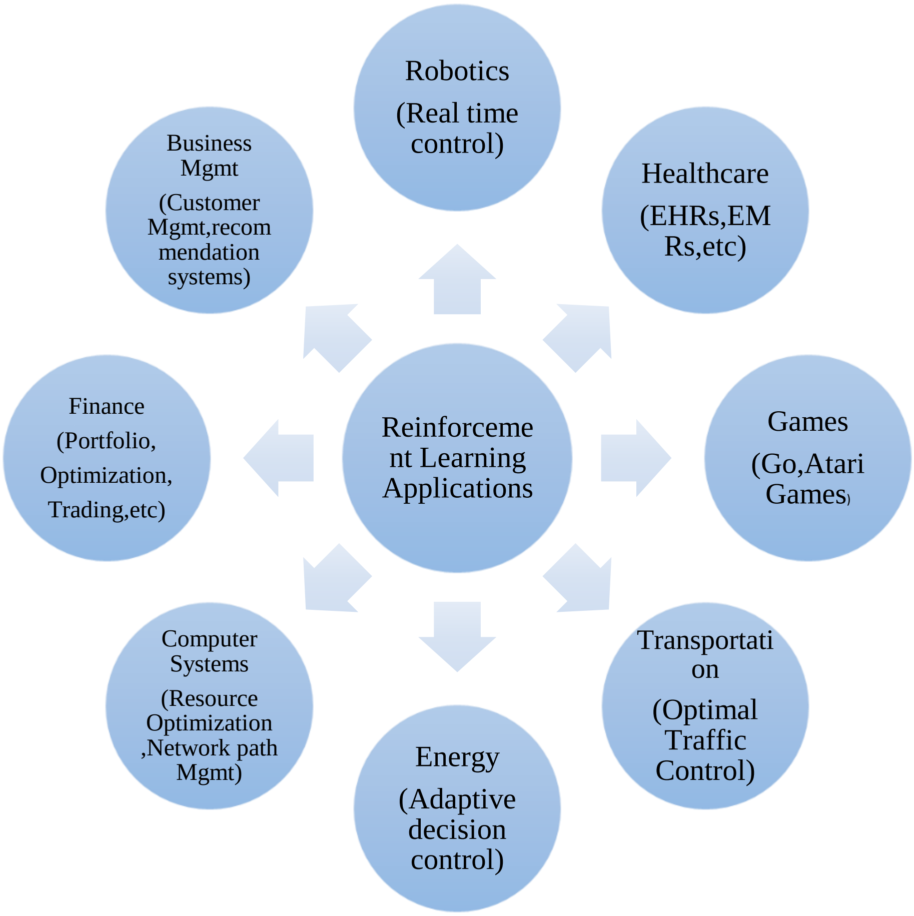

Figure 1. Applications for RL [32].

Table 1. Related works contribution.

No	Survey paper	Technique	Field	Contribution
1	[33]	Deep Reinforcement Learning	Medical	This study covers the fundamentals of RL and provides an overview of the several types of DRL models. Researchers have created algorithms for radiation therapy planning optimization and medical picture interpretation.
2	[34]	Inverse Reinforcement Learning	Represents an area of the literature	This work provides an extensive overview of the IRL literature.

			that needs more development	The survey delineates the distinctions between IRL and two analogous techniques, namely inverse optimal control and apprenticeship learning.
3	[35]	Reinforcement Learning	Education	The paper addresses concerns regarding the efficacy and prospects of RL in education. The paper offers a comprehensive review of the many methods and uses of RL in this field.
4	[36]	Reinforcement Learning	Exploration	The survey provides a thorough synopsis of current exploration methodologies. The survey addresses open issues to offer important avenues for further research.
5	[37]	Deep Reinforcement Learning	Economics	DRL outperforms traditional algorithms and is more efficient when faced with actual economic problems in the presence of risk characteristics and ever-increasing uncertainties.
6	[38]	Reinforcement Learning	Mathematics	This overview reveals the mathematical foundations to help readers comprehend the main ideas and apply them to their study.
7	[5]	Deep Reinforcement Learning	AI	The study addresses key DRL techniques, such as the asynchronous advantage actor-critic, trust region policy optimization, and deep Q-network (DQN).
8	[39]	Hierarchical Reinforcement Learning	Learning hierarchical policies, subtask discovery, transfer learning, and multi-agent learning	This paper offers an overview of the various HRL methodologies and discusses the difficulties in applying HRL to learn hierarchical policies, subtask discovery, transfer learning, and multi-agent learning.
9	[40]	Reinforcement Learning	Machine learning	This paper demonstrates and elucidates RL interpretation techniques, categorizes the metrics used, and explains how these metrics are employed to understand the inner workings of RL models.
10	[41]	Reinforcement Learning	Robotic applications	This survey provides an overview of the use of reinforcement algorithms in robotic research.

2 | Background

2.1 | Convolutional Neural Network

Three forms of learning technologies exist for RL, and one of the most significant machine learning techniques [42]: is non-supervised learning, supervised learning, and RL. Compared to supervised and non-supervised learning, RL is an online learning method.

Fundamentally, RL is an interactive machine learning paradigm in which an agent engages with the environment, gains experience, and applies that experience. To enhance its guidelines. We see a significant gap in RL's capacity to generalize when compared to other ML paradigms; RL has mostly succeeded in limited and very narrow domains [43, 44].

2.2 | Markov Decision Processes

Mathematical models called Markov Decision Processes (MDPs) [45] are used to describe how an agent interacts with its surroundings. An MDP is officially represented as a tuple of five items (S, A, P, R, γ), where S represents the set of potential states or the state space. The action space, or the collection of potential actions, is represented by A. $P: S \times A \times S \rightarrow [0,1]$ shows the likelihood of changing from one state to another given a specific activity. $R = S \times A \times S \rightarrow R$, the reward function is denoted by R, whereas the discount factor γ establishes the significance of upcoming rewards, $\gamma \epsilon [0,1]$. discrete-time steps are used by the agent to interact with its surroundings.$t = 0,1,2, \ldots.;$

The agent obtains a representation of the environmental state $S_t \in S$, at each time step t. It then acts $A_t \in A$, advances to the next stage S_{t+1}, and is rewarded with a scalar value$R_{t+1} \in R$. This is the conventional RL structure shown in Figure 2.

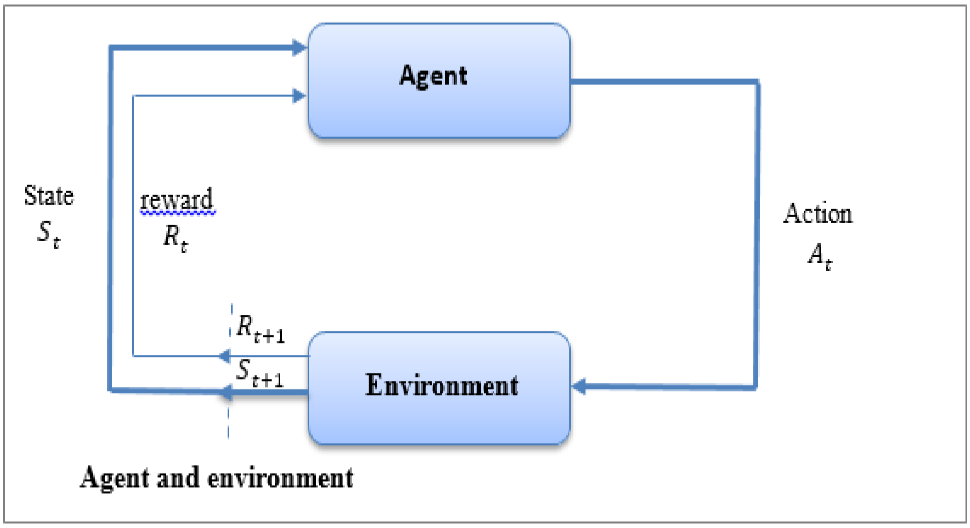

Figure 2. Interaction between the agent and the environment [46].

A policy is the behavior of the agent that associates states with actions, where $\pi: S \times A$ is the probability of doing an action $a \in A$ given a state s is $\pi(s|a) = \Pr(A_t = a|S_t = s)$ .The agent's objective is to maximize the return, or expected cumulative discounted reward, which is represented by the symbol Gt:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \quad (1)$$

Where:

G_t : The return at time t. This is the total discounted reward from time t onward.

γ: This is the discount factor, which is a number between 0 and 1. It determines the present value of future rewards.

R_{t+k+1}: The reward received at time $t + k + 1$. this represents the reward the agent gets at each future time step.

$\gamma \in [0,1]$ The typical value and γ is the discount factor. The optimum policy, denoted by π^*, is the behavior that maximizes reward over time by choosing the best course of action in each stage.

Additionally, model-based and model-free algorithms are subsets of RL-based recommendation models. While the majority of recommendation models now in use employ model-free algorithms, a small number of them use model-based methods, as seen in Figure 3.

Model-free and model-based RL algorithms may be classified into two primary types based on whether the agent uses an environment dynamics model that can be learned or given the reward function, R, and the transition function, P, are described by the model. The model-based techniques fall into two categories: those that employ a predefined model (the agent may access the reward function and transition models) and those that teach the agent the environment model [47].

With the latter method, the agent gains knowledge about a model that it utilizes to enhance policies. By acting, the agent can gather samples from the surrounding area. From those examples, rewards and state transitions may be anticipated using supervised learning. The environment model may be directly utilized with planning techniques. Instead of attempting to create a model of the environment, the agent in the model-free method interacts with the environment to choose the best course of action via trial and error. Model-free approaches are simpler to put into practice than model-based approaches. When creating an accurate enough model proves to be challenging, these approaches may prove to be more beneficial than more intricate ones [46].

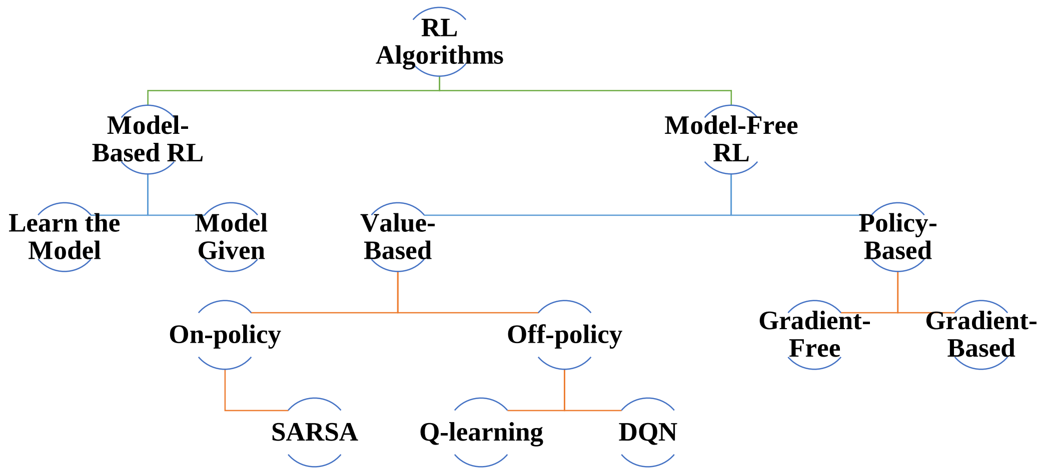

Figure 3. Classification of RL algorithms.

- **Model-free RL**

  Instead of using the action and observation data to train a transition model, model-free approaches try to give a value directly to a state or a state-action combination. The action value function, also known as the Q-function, or an ensemble of them, is trained using the offline RL techniques that are examined below. Using the Bellman optimality operator, Q-learning techniques maximize the Q-function [48].

  Using two deep neural networks (DNNs), DDQN is an enhanced version of the classic Q-learning technique that addresses the problems of dimensional explosion and Q-value overestimation. The agent searches each state for the action that produces the highest Q-value. The predicted reward attained for doing a certain action at a given state s is known as the Q-value, also known as the state-action value [49].

- **Model-based RL**

  The state-of-the-art model-based RL technique, the model-based policy optimization framework (MBPO), is used. It was first developed by Janner et al. [50] One framework that integrates learning and planning is called MBPO, which is a Dyna-style algorithm. When employing a model-based approach, states and behaviors are predicted using an environment model, whereas a policy

  Finding a strategy that maximizes the expected benefit is the goal of the optimization technique. MBPO can offer a more effective method of resolving control issues and demonstrates strong control results by combining these two techniques [51].

3 | Comparison of the Applications of RL in Social Sciences

In this section, we will discuss RL social sciences.

3.1 | Economics and Finance

Charpentier et al. [52] suggested RL techniques in the fields of finance, game theory, operation research, and economics. RL algorithms explain how, with repeated experience, an agent may figure out the best course of action in a sequential decision-making process. Zheng et al. [53] suggest using AI economists to create DRL economic policies that are optimum to solve problems with counterfactual data, behavioral models, policy evaluation, and behavioral reactions. Bacoyannis et al. [54] described the peculiarities of machine learning and neural information processing in the context of quantitative finance. Boukas et al. [55] suggested a unique modeling framework for energy storage's strategic involvement in the European continuous intraday market, where transactions take place via a centralized order book.

3.2 | Gaming

Castañón et al. [56] Proposed an agent-based model based on the Bush–Mosteller RL algorithm is proposed. After playing rounds of the Dictator Game, agents update their aspirations (and, consequently, their future cooperative behavior) in response to stimuli derived from empirical and normative expectations. Zhao et al. [57] Introducing AlphaHoldem and using end-to-end RL (without CFR) to obtain excellent performance. DeepStack, Libratus, and Alpha-Holdem are algorithms designed for two-player zero-sum games with incomplete information, which are a challenging class of issues. Wurman et al. [58] created a car racing agency and won over the top e-sports drivers globally.

3.3 | Psychology

Mnih et al. [43] suggested Leveraging current developments in deep neural network training to create a unique artificial agent known as a deep Q-network that uses end-to-end RL to learn effective policies directly from high-dimensional sensory inputs. Doroudi et al. [59] explain how RL is used for instructional sequencing and demonstrate how concepts and theories from learning sciences and cognitive psychology might enhance performance.

3.4 | Sociology

Jaques et al. [60] suggested a unified method for Multi-Agent RL (MARL) that rewards actors for their causal impact on the behaviors of other agents to achieve coordination and communication. Counterfactual reasoning is used to evaluate causal influence. An agent simulates alternative actions it may have done at each time step and calculates the impact of those actions on other agents' behavior.

Weltz et al. [61] in this paper, RL is a perfect model for many difficult decision issues that come up in public health, such as allocating resources during a pandemic, testing or monitoring, and adaptive sampling for populations that are concealed.

3.5 | Political Science

Schulz et al. [62] suggested using RL as a cohesive framework for analyzing political thought. RL explains the algorithmic navigation of complicated and unpredictable environments such as politics by agents. Using this computational perspective, they delineate three pathways leading to political disparities, which originate from variations in agents' perceptions of an issue, the mental processes utilized to address the issue or the context of accessible environmental data. A computational perspective on political mental illnesses provides more accuracy in determining their origins, effects, and treatments.

3.6 | Medical

Zhu et al. [63] Suggested diagnostic strategy learning in this study, and a novel framework including three components is proposed to learn a diagnostic strategy with restricted features. Gottesman et al. [64] examined interpretable RL by emphasizing significant transitions and using it with data from intensive care units (ICUs) and medical simulations. Capobianco et al. [65] examined how to best implement mitigation strategies while taking hospital capacity and the economy into account. Colas et al. [66] suggested using bandits algorithms in Greece for COVID-19 testing.

3.7 | Education

Fu et al. [67] suggested that teaching and learning quality are negatively impacted by the incorrect identification of students' learning skills is addressed by employing digital smart classrooms that support the learning features of the students because social factors and the students' behavior have an impact on learning efficiency. Oudeyer et al. [68] elucidate how kinds of mechanisms of interest may be modeled within the framework of computational RL. Cai et al. [69] provided a proposed instructional conversational agent that combines rules and contextual bandits to provide practice questions, explanations of arithmetic topics, and personalized feedback. Singla et al. [70] planned a workshop as a means of fostering a sense of community among scholars and professionals engaged in the general fields of education (ED) and RL. The purpose of this article is to give a summary of the key research directions in the field of RL for ED and to give an overview of the workshop events.

There is a comparison of RLs in different social science applications such as economics, gaming, Political Science, and Education are presented in Table 2.

Table 2. Comparison of the applications of RL in social sciences.

No	Reference	Application	Methodology	Contribution
1	[52]	Economics and Finance	Reinforcement Learning	This paper presents RL techniques and their applications in economics and finance.
2	[53]	Economics and Finance	Reinforcement Learning	They encourage AI economists to create the best possible economic policy using deep RL. Researchers approach this work as an optimization problem, where cutting-edge machine learning techniques like DRL are quite beneficial.
3	[54]	Economics and Finance	Reinforcement Learning	They discuss the peculiarities and challenges of data-driven learning in online trading.
4	[55]	Economics and Finance	Deep Reinforcement Learning	This paper proposes a novel modeling framework for examining ongoing intraday market bidding.
5	[56]	Gaming (dictator game experiment)	Agent-based Model and Reinforcement Learning	They explain how social standards influence RL agents
6	[57]	Gaming	Reinforcement Learning	The suggested system uses a pseudo-Siamese architecture to compare the learned model with several previous iterations to directly learn from the input state information to the output actions.
7	[58]	Gaming	Deep Reinforcement Learning	They combine state-of-the-art, model-free, DRL algorithms to build an integrated control policy that combines remarkable speed and strategies.

8	[43]	psychology	Deep Reinforcement Learning	They use recent advances in training deep neural networks to develop a novel artificial agent that Bridges the gap between actions and high-dimensional sensory inputs
9	[59]	psychology	Reinforcement Learning	They discover that situations where RL has been limited by concepts and theories from cognitive psychology and the learning sciences have had the most success with it.
10	[61]	public health	Reinforcement Learning	This paper presents important concepts in RL and points out potential applications and obstacles in public health RL.
11	[60]	Sociology	Multi-Agent Reinforcement Learning (MARL)	They suggest a unified method for accomplishing cooperation and communication by rewarding actors for their causal impact on the behaviors of other agents.
12	[62]	Political Science	Reinforcement Learning	They present a Cohesive framework for analyzing political thought
13	[63]	Medical	Deep Reinforcement Learning	This method offers individualized diagnostic strategies and produces better diagnoses with fewer characteristics.
14	[64]	Medical	Reinforcement Learning	The authors of this work have developed a method that might function as a hybrid human-AI system.
15	[65]	Medical	Reinforcement Learning	This paper studies how to optimize mitigation approaches that minimize the impact on the economy without overstuffing hospital capacity using Bayesian inference and RL.
16	[66]	Medical	Deep Reinforcement Learning	They demonstrate how to apply a Susceptible-Exposed-Infectious-Removed (SEIR) model for COVID-19 to determine the best policies for dynamical on-lockdown control while optimizing the death toll and economic recession.
17	[67]	Education	Reinforcement Learning	This work uses RL to generate intelligent and comfortable learning.
18	[68]	Education	Reinforcement Learning	They discuss the learning progress (LP) theory, which suggests that learning and curiosity create a positive feedback loop. They describe robot studies that demonstrate how LP-driven exploration and attention can self-organize a curriculum for developmental learning, effectively scaffolding the acquisition of various skills and tasks.
19	[69]	Education	Reinforcement Learning	This work shows that conversational agents hold great potential as a supplement to current web-based resources for math instruction.
20	[70]	Education	Reinforcement Learning	They planned the workshop to foster a sense of community among scholars and professionals in the fields of education (ED) and reinforcement learning (RL). This article summarizes key research directions in the field of RL for education and provides an overview of the workshop events.

4 | Future Work and Challenges

RL's tremendous success in the recent past has been on a wide range of problems, in areas ranging from robotics to playing games. However, broad and challenging areas open to further research, exist. This survey examines future directions and challenges in the area by describing important areas in which improvements must be made if RL applications are to become more general and robust. RL a subfield of machine learning dedicated to control problems, is emerging as a potentially revolutionary approach to building controls. Because RL is data-driven, users may be able to avoid the laborious process of creating and fine-tuning a detailed model, which is necessary for MPC. Moreover, RL may be able to take advantage of the recent and swift advancements in the field of machine learning, such as deep learning and feature encoding, to improve control decisions [71].

In the energy system domain, the state of the art clearly shows that RL can be used effectively for a variety of control problems. More importantly, it can be used to solve complex problems, like those in sector coupling, which can significantly help with energy transition and climate change mitigation. It would be interesting to explore the potential of RL beyond just controlling energy flows. While RL has been successfully applied with supervised and unsupervised learning in other sectors, there aren't many examples in the energy system domain [72]. Also, one of the challenges is access to high-quality, ethically sourced social data remains crucial for training and validating RL models.

RL has come a long way, but there are still a lot of problems. Common problems include sample efficiency, credit assignment, exploration vs. exploitation, and representation. Problems arise when using value function techniques with function approximation. DRL has a reproducibility problem, meaning that many hyperparameters such as reward size and network design, random seeds and trials, settings, and codebases [73] can affect the outcome of experiments. Problems with reward specifications can arise, and a reward function might not accurately reflect the designer's goal. Issues like the expressivity of Markov reward [74] and delusional bias [75] are still being recognized and addressed by researchers and practitioners.

Utilizing massive volumes of unlabeled data with unsupervised RL techniques is a potential future approach for the profession [76]. Labeling big datasets with rewards may often be expensive, particularly if human oversight is needed. Using different unlabeled data in an easy-to-use but efficient way is still an unsolved issue. Yu et al. [77] demonstrate how a limited quantity of high-quality labeled data along with a large number of inferior unlabeled data may be used to develop successful strategies. Similar findings are presented by Kumar et al. [78] when contrasting offline RL with BC techniques. Yarats et al. [79] demonstrate how to leverage a variety of unlabeled datasets with downstream reward relabeling to improve the effectiveness of standard off-policy RL techniques [80] in offline environments.

Also, we illustrate a few benefits of using offline RL over online RL for a particular application using current instances. Emerson et al. [30] employed offline RL in the healthcare domain to create a policy that determines the ideal insulin dosage to sustain blood glucose levels within a healthy range. They contend that online role-playing is simply too erratic to control blood sugar levels and may push patients beyond their safe threshold. Zhan et al. [81]offer a model-based offline RL approach for energy management that maximizes the thermal power generating units' (TPGUs') combustion control strategy. Large volumes of historical TPGU data combined with low-fidelity simulation data allow them to develop a safety-constrained strategy that significantly outperforms BC. Using the available data to create a policy in this instance was significantly less costly and time-consuming than doing it interactively. Ultimately, Verma et al. [31] propose training a task-driven conversation system with offline RL. Agent known as CHatbot AI, or CHAI.

RL's purpose is to identify an optimal policy - a mapping from world conditions to a set of behaviors - that maximizes cumulative reward, which is a long-term strategy. Exploring may be suboptimal in the near term, but it may result in excellent outcomes in the long run. Many optimal control problems, which have been popular in economics for over four decades, can be expressed in the RL framework, and economists can use recent advances in computational science, particularly deep learning algorithms, to solve complex behavioral problems.

When data scientists are well-versed in RL, they may enhance their models and raise the bar for performance. More significantly, RL techniques can frequently outperform human supervisors while offering scientists and researchers fresh viewpoints and a greater comprehension of these difficulties. It is our aim that readers will be able to draw parallels between these studies, get a deeper understanding of RL concepts, and use RL in their future research.

The future of RL in social sciences lies in developing robust, interpretable, and ethically sound models. This will allow us to illuminate complex social phenomena, ultimately informing better policies and interventions. This survey provides a starting point for further research, and collaboration between social scientists and RL experts is the key to unlocking the true potential of this powerful tool for understanding the social world around us.

5 | Conclusion

Finally, the primary goal of this study was to give both novice and seasoned researchers in the field a comprehensive grasp of the use of RL in the social sciences, hence directing future studies and advancements in this subject. RL has been successfully used in several significant real-world contexts. The purpose of this survey is to introduce the Markov Decision Process. Additionally, we provide an overview of the literature on the use of RL in a range of disciplines, such as political science, psychology, gaming, economics and finance, medicine, and education. We present a thorough introduction to RL in this survey. First, we offer a categorization of every RL algorithm. Lastly, we offer our thoughts on the unresolved issues in the discipline, along with some encouraging research avenues for the future.

Acknowledgments

The author is grateful to the editorial and reviewers, as well as the correspondent author, who offered assistance in the form of advice, assessment, and checking during the study period.

Author Contributions

All authors contributed equally to this work.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] D. Bertsekas, Reinforcement learning and optimal control vol. 1: Athena Scientific, 2019.
[2] H. Dong, H. Dong, Z. Ding, S. Zhang, and T. Chang, Deep Reinforcement Learning: Springer, 2020.
[3] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," Nature Machine Intelligence, vol. 1, pp. 133-143, 2019.
[4] H. Li, D. Liu, and D. Wang, "Manifold regularized reinforcement learning," IEEE transactions on neural networks and learning systems, vol. 29, pp. 932-943, 2017.
[5] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," IEEE Signal Processing Magazine, vol. 34, pp. 26-38, 2017.
[6] G. Zheng, F. Zhang, Z. Zheng, Y. Xiang, N. J. Yuan, X. Xie, et al., "DRN: A deep reinforcement learning framework for news recommendation," in Proceedings of the 2018 world wide web conference, 2018, pp. 167-176.
[7] D. Zha, L. Feng, B. Bhushanam, D. Choudhary, J. Nie, Y. Tian, et al., "Autoshard: Automated embedding table sharding for recommender systems," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 4461-4471.
[8] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," The International Journal of Robotics Research, vol. 32, pp. 1238-1274, 2013.
[9] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, et al., "Human-level performance in 3D multiplayer games with population-based reinforcement learning," Science, vol. 364, pp. 859-865, 2019.
[10] L. Zou, L. Xia, Y. Gu, X. Zhao, W. Liu, J. X. Huang, et al., "Neural interactive collaborative filtering," in Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, 2020, pp. 749-758.
[11] Q. Liu, S. Tong, C. Liu, H. Zhao, E. Chen, H. Ma, et al., "Exploiting cognitive structure for adaptive learning," in Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 627-635.
[12] S. Ji, Z. Wang, T. Li, and Y. Zheng, "Spatio-temporal feature fusion for dynamic taxi route recommendation via deep reinforcement learning," Knowledge-Based Systems, vol. 205, p. 106302, 2020.
[13] H. Lee, D. Hwang, K. Min, and J. Choo, "Towards validating long-term user feedbacks in interactive recommendation systems," in Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2607-2611.
[14] K. Wang, Z. Zou, Q. Deng, J. Tao, R. Wu, C. Fan, et al., "Reinforcement learning with a disentangled universal value function for item recommendation," in Proceedings of the AAAI conference on artificial intelligence, 2021, pp. 4427-4435.
[15] M. L. Puterman, Markov decision processes: discrete stochastic dynamic programming: John Wiley & Sons, 2014.
[16] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in International conference on machine learning, 2014, pp. 387-395.
[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, et al., "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.
[18] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," Journal of Intelligent & Robotic Systems, vol. 86, pp. 153-173, 2017.
[19] N. R. Ke, A. Singh, A. Touati, A. Goyal, Y. Bengio, D. Parikh, et al., "Learning dynamics model in reinforcement learning by incorporating the long term future," arXiv preprint arXiv:1903.01599, 2019.
[20] Z. Ding, Y. Huang, H. Yuan, and H. Dong, "Introduction to reinforcement learning," Deep reinforcement learning: fundamentals, research and applications, pp. 47-123, 2020.
[21] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in Proceedings of the fifth international conference on Knowledge capture, 2009, pp. 9-16.
[22] C. Celemin and J. Ruiz-del-Solar, "An interactive framework for learning continuous actions policies based on corrective feedback," Journal of Intelligent & Robotic Systems, vol. 95, pp. 77-97, 2019.
[23] A. L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners," Artificial Intelligence, vol. 172, pp. 716-737, 2008.
[24] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-centered reinforcement learning: A survey," IEEE Transactions on Human-Machine Systems, vol. 49, pp. 337-349, 2019.
[25] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," arXiv preprint arXiv:1904.12901, 2019.
[26] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," arXiv preprint arXiv:2005.01643, 2020.
[27] Y. Fu, W. Di, and B. Boulet, "Batch reinforcement learning in the real world: A survey," in Offline RL Workshop, NeuroIPS, 2020.
[28] L. Zhang, R. Zhang, T. Wu, R. Weng, M. Han, and Y. Zhao, "Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles," IEEE transactions on neural networks and learning systems, vol. 32, pp. 5435-5444, 2021.
[29] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng, "Reinforcement learning for clinical decision support in critical care: comprehensive review," Journal of medical Internet research, vol. 22, p. e18477, 2020.

[30] H. Emerson, M. Guy, and R. McConville, "Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes," Journal of Biomedical Informatics, vol. 142, p. 104376, 2023.

[31] S. Verma, J. Fu, M. Yang, and S. Levine, "Chai: A chatbot ai for task-oriented dialogue with offline reinforcement learning," arXiv preprint arXiv:2204.08426, 2022.

[32] R. Nannapaneni, "Optimal path routing using reinforcement learning," DELLTECHNOLOY giES, pp. 1-25, 2020.

[33] L. Xu, S. Zhu, and N. Wen, "Deep reinforcement learning and its applications in medical imaging and radiation therapy: a survey," Physics in Medicine & Biology, vol. 67, p. 22TR02, 2022.

[34] S. Adams, T. Cody, and P. A. Beling, "A survey of inverse reinforcement learning," Artificial Intelligence Review, vol. 55, pp. 4307-4346, 2022.

[35] B. Fahad Mon, A. Wasfi, M. Hayajneh, A. Slim, and N. Abu Ali, "Reinforcement Learning in Education: A Literature Review," in Informatics, 2023, p. 74.

[36] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: A survey," Information Fusion, vol. 85, pp. 1-22, 2022.

[37] A. Mosavi, Y. Faghan, P. Ghamisi, P. Duan, S. F. Ardabili, E. Salwana, et al., "Comprehensive review of deep reinforcement learning methods and applications in economics," Mathematics, vol. 8, p. 1640, 2020.

[38] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," INFORMS Journal on Computing, vol. 21, pp. 178-192, 2009.

[39] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," ACM Computing Surveys (CSUR), vol. 54, pp. 1-35, 2021.

[40] A. Alharin, T.-N. Doan, and M. Sartipi, "Reinforcement learning interpretation methods: A survey," IEEE Access, vol. 8, pp. 171058-171077, 2020.

[41] B. Singh, R. Kumar, and V. P. Singh, "Reinforcement learning in robotic applications: a comprehensive survey," Artificial Intelligence Review, vol. 55, pp. 945-990, 2022.

[42] R. R. Bush and F. Mosteller, "Stochastic Models for," Learning, 1955.

[43] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, et al., "Human-level control through deep reinforcement learning," nature, vol. 518, pp. 529-533, 2015.

[44] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, et al., "Mastering the game of Go with deep neural networks and tree search," nature, vol. 529, pp. 484-489, 2016.

[45] N. Akalin and A. Loutfi, "Reinforcement learning approaches in social robotics," Sensors, vol. 21, p. 1292, 2021.

[46] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction: MIT press, 2018.

[47] H. Zhang and T. Yu, "Taxonomy of reinforcement learning algorithms," Deep reinforcement learning: Fundamentals, research and applications, pp. 125-133, 2020.

[48] P. Swazinna, S. Udluft, D. Hein, and T. Runkler, "Comparing Model-free and Model-based Algorithms for Offline Reinforcement Learning," IFAC-PapersOnLine, vol. 55, pp. 19-26, 2022/01/01/ 2022.

[49] M. Esrafilian-Najafabadi and F. Haghighat, "Towards self-learning control of HVAC systems with the consideration of dynamic occupancy patterns: Application of model-free deep reinforcement learning," Building and Environment, vol. 226, p. 109747, 2022.

[50] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," Advances in neural information processing systems, vol. 32, 2019.

[51] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, et al., "Mopo: Model-based offline policy optimization," Advances in Neural Information Processing Systems, vol. 33, pp. 14129-14142, 2020.

[52] A. Charpentier, R. Elie, and C. Remlinger, "Reinforcement learning in economics and finance," Computational Economics, pp. 1-38, 2021.

[53] A. Trott, S. Srinivasa, D. van der Wal, S. Haneuse, and S. Zheng, "Building a foundation for data-driven, interpretable, and robust policy design using the ai economist," arXiv preprint arXiv:2108.02904, 2021.

[54] V. Bacoyannis, V. Glukhov, T. Jin, J. Kochems, and D. R. Song, "Idiosyncrasies and challenges of data driven learning in electronic trading," arXiv preprint arXiv:1811.09549, 2018.

[55] I. Boukas, D. Ernst, T. Théate, A. Bolland, A. Huynen, M. Buchwald, et al., "A deep reinforcement learning framework for continuous intraday market bidding," Machine Learning, vol. 110, pp. 2335-2387, 2021.

[56] R. Castañón, F. A. Campos, J. Villar, and A. Sánchez, "A reinforcement learning approach to explore the role of social expectations in altruistic behavior," Scientific Reports, vol. 13, p. 1717, 2023.

[57] E. Zhao, R. Yan, J. Li, K. Li, and J. Xing, "Alphaholdem: High-performance artificial intelligence for heads-up no-limit poker via end-to-end reinforcement learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 4689-4697.

[58] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, et al., "Outracing champion Gran Turismo drivers with deep reinforcement learning," Nature, vol. 602, pp. 223-228, 2022.

[59] S. Doroudi, V. Aleven, and E. Brunskill, "Where's the reward? a review of reinforcement learning for instructional sequencing," International Journal of Artificial Intelligence in Education, vol. 29, pp. 568-620, 2019.

[60] N. Jaques, A. Lazaridou, E. Hughes, C. Gulcehre, P. Ortega, D. Strouse, et al., "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in International conference on machine learning, 2019, pp. 3040-3049.

[61] J. Weltz, A. Volfovsky, and E. B. Laber, "Reinforcement learning methods in public health," Clinical therapeutics, vol. 44, pp. 139-154, 2022.
[62] L. Schulz and R. Bhui, "Political reinforcement learners," Trends in Cognitive Sciences, vol. 28, pp. 210-222, 2024/03/01/ 2024.
[63] M. Zhu and H. Zhu, "Learning a diagnostic strategy on medical data with deep reinforcement learning," IEEE Access, vol. 9, pp. 84122-84133, 2021.
[64] O. Gottesman, J. Futoma, Y. Liu, S. Parbhoo, L. Celi, E. Brunskill, et al., "Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions," in International Conference on Machine Learning, 2020, pp. 3658-3667.
[65] R. Capobianco, V. Kompella, J. Ault, G. Sharon, S. Jong, S. Fox, et al., "Agent-based Markov modeling for improved COVID-19 mitigation policies," Journal of Artificial Intelligence Research, vol. 71, pp. 953-992, 2021.
[66] C. Colas, B. Hejblum, S. Rouillon, R. Thiébaut, P.-Y. Oudeyer, C. Moulin-Frier, et al., "Epidemioptim: A toolbox for the optimization of control policies in epidemiological models," Journal of Artificial Intelligence Research, vol. 71, pp. 479-519, 2021.
[67] S. Fu, "A reinforcement learning-based smart educational environment for higher education," International Journal of e-Collaboration (IJeC), vol. 19, pp. 1-17, 2022.
[68] P.-Y. Oudeyer, J. Gottlieb, and M. Lopes, "Intrinsic motivation, curiosity, and learning: Theory and applications in educational technologies," Progress in brain research, vol. 229, pp. 257-284, 2016.
[69] W. Cai, J. Grossman, Z. J. Lin, H. Sheng, J. T.-Z. Wei, J. J. Williams, et al., "Bandit algorithms to personalize educational chatbots," Machine Learning, vol. 110, pp. 2389-2418, 2021.
[70] A. Singla, A. N. Rafferty, G. Radanovic, and N. T. Heffernan, "Reinforcement learning for education: Opportunities and challenges," arXiv preprint arXiv:2107.08828, 2021.
[71] Z. Wang and T. Hong, "Reinforcement learning for building controls: The opportunities and challenges," Applied Energy, vol. 269, p. 115036, 2020/07/01/ 2020.
[72] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," Renewable and Sustainable Energy Reviews, vol. 137, p. 110618, 2021/03/01/ 2021.
[73] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in Proceedings of the AAAI conference on artificial intelligence, 2018.
[74] D. Abel, W. Dabney, A. Harutyunyan, M. K. Ho, M. Littman, D. Precup, et al., "On the expressivity of markov reward," Advances in Neural Information Processing Systems, vol. 34, pp. 7799-7812, 2021.
[75] T. Lu, D. Schuurmans, and C. Boutilier, "Non-delusional Q-learning and value-iteration," Advances in neural information processing systems, vol. 31, 2018.
[76] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, et al., "Scaling data-driven robotics with reward sketching and batch reinforcement learning," arXiv preprint arXiv:1909.12200, 2019.
[77] T. Yu, A. Kumar, Y. Chebotar, K. Hausman, C. Finn, and S. Levine, "How to leverage unlabeled data in offline reinforcement learning," in International Conference on Machine Learning, 2022, pp. 25611-25635.
[78] A. Kumar, J. Hong, A. Singh, and S. Levine, "Should i run offline reinforcement learning or behavioral cloning?," in International Conference on Learning Representations, 2021.
[79] D. Yarats, D. Brandfonbrener, H. Liu, M. Laskin, P. Abbeel, A. Lazaric, et al., "Don't change the algorithm, change the data: Exploratory data for offline reinforcement learning," arXiv preprint arXiv:2201.13425, 2022.
[80] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in International conference on machine learning, 2018, pp. 1587-1596.
[81] X. Zhan, H. Xu, Y. Zhang, X. Zhu, H. Yin, and Y. Zheng, "Deepthermal: Combustion optimization for thermal power generating units using offline reinforcement learning," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022, pp. 4680-4688.