

International Journal of Advances and Applications in Convergent Studies & Transdisciplinary
(IJ-AACST)

Bridging Disciplines, Bridging Knowledge Aligned
with the Vision of NEP 2020 and UGC Guidelines



Website: <https://aacstij.org>

||

Email: journaleditor@aacstij.org

Reinforcement Learning: Emerging Trends and Future Research Directions

DOI: <https://doi.org/10.5281/zenodo.23162406>

Yash D. Turkey

Research Student, M.Sc. Computer Science, R.A. Arts, Shri M.K. Commerce and Shri S.R. Rath
Science College, Washim - 444505, Maharashtra, India.

yasht.1702@gmail.com

ORCID: 0009-0001-5685-5089

ABSTRACT

Reinforcement learning (RL) is a small solved part of sequential decision-making theory, and has emerged over the last ten years as one of the most active and productive fields of research in the field of artificial intelligence (AI). It has propelled advancements across various fields, including game playing, robot control, and the recently explored application of aligning large language models with human goals. This paper provides a brief overview of the research field in a structured way. First, it presents the formalism of the Markov decision process and the main classes of algorithms developed on top of it, and subsequently analyzes six research directions that are driving the field today, namely deep reinforcement learning, model-based reinforcement learning, offline reinforcement learning, multi-agent reinforcement learning, reinforcement learning from human feedback for language-model alignment, and safe and robust reinforcement learning. It then summarizes some typical application areas (robotics, autonomous driving, healthcare, communication networks, and games) and introduces some open problems: sample efficiency, reward design, generalization, and safety verification. It ends with the presentation of future research opportunities such as hybrid models-based/model-free architectures, online/offline training pipelines, scalable ways of human supervision, and RL approaches for future communication networks.

Keywords: Reinforcement learning, deep reinforcement learning, offline reinforcement learning, multi-agent systems, reinforcement learning from human feedback, safe reinforcement learning, Markov decision process, autonomous systems.

1. INTRODUCTION

Reinforcement learning (RL) is a subset of machine learning that involves one or more artificial agents that interact with an environment and learn to maximize cumulative reward over time by trial and error. In contrast to supervised learning, where a model is built using examples of the correct answer, an RL agent receives only reflective feedback from the environment in the form of scalar reward signals generated by its interactions with the environment. This trial-and-error learning process can be structured as a Markov decision process (MDP) [1]. With function approximation using deep neural networks, the framework has yielded some of the most impressive achievements to date in the application of AI to the past decade of Atari games, such as agents playing many Atari games at superhuman level directly from raw pixel input [2], and agents that beat top human players of Go [3].

Over the last decade or so, a new line of work that has been called deep reinforcement learning (DRL) has resurfaced in the spotlight of RL. Deep networks enable the use of value functions and policies in high-dimensional spaces, not just in small, discrete spaces, allowing them to solve problems with a rich sensory input, continuous control, and long planning horizon. Concurrently, the number of different variants of algorithms, standardized benchmarks, and application areas has increased greatly; and various survey articles have tried to classify and compare these [4].

Now a second wave of change is taking place in the field, with at least four trends playing out. One is that trial and error learning typically is impractical, expensive or even unsafe in practice, and as such, offline (batch) reinforcement learning has gained significant traction, where a policy is learned offline using previously gathered interaction data without relying on new interaction [5]. Second, RL has risen to prominence beyond its control-theoretic roots, since the advent of large language models, for example in the alignment of large language models with human preferences, via reinforcement learning from human feedback (RLHF) and its variants [6]. Third, in the safety-critical areas, including transportation, industrial control, and health care, where the use of RL is becoming more prevalent, safety and robustness has emerged as formal properties and a main research interest. Fourth, there are many practical problems in reality that are inherently multi-agent, and that necessitates a multi-agent reinforcement learning where multiple agents need to learn in parallel and cooperate with each other, which is still a hot topic in the field of reinforcement learning [8].

2. CONTRIBUTION OF PAPER

I. This paper gives a structured overview of recent reinforcement learning research. It starts from the Markov decision process formalism and the main algorithm families (value-based, policy-based and actor-critic, plus model-free versus model-based methods). It then organizes current research into six themes: deep RL, model-based RL, offline RL, multi-agent RL, RLHF for language-model alignment,

and safe and robust RL. It also shows how these themes overlap, for example offline pre-training followed by online fine-tuning, or safety constraints in multi-agent systems.

II. The paper summarizes how RL is applied in robotics, autonomous driving, healthcare, communication networks and IoT, and games. It gives particular attention to adaptive routing, resource allocation and 6G network management. It also identifies the open problems of sample efficiency, reward specification, generalization and safety verification. It closes with future directions: integration with foundation models, hybrid model-based/model-free approaches, offline-to-online learning, formal verification and shielding, scalable human oversight, and learning from expert demonstrations.

3. FOUNDATIONAL CONCEPTS

The RL problem is usually represented as a MDP with a set of states, a set of actions, a transition distribution depending on the action, a reward function (which is usually unknown to the agent), and a discount factor applied to future rewards as compared to immediate rewards. The agent observes the environment state, chooses an action based on its policy, is awarded a scalar reward and transitions to the next state according to the environment dynamics. The goal of the agent is to locate a policy that will generate the greatest sum of discounted future rewards: the return [9].

There are two general classes of algorithms that are designed to achieve this goal. Value-based methods learn the expected return of a single state or state–action pair, either by learning a value function or an action-value function, and choose the action with the highest value by selecting the action that is associated with the highest value; examples include Q-learning and its deep variant, the Deep Q-Network [2]. In contrast, policy-based methods parameterize the policy directly, and tune the parameters according to gradient ascent of the expected returns. A third family, actor-critic methods, is a mix of the two, in which the actor learns the policy, and the critic learns value functions to decrease the variance in the policy-gradient estimate. The clipped surrogate objective is a widely used actor–critic algorithm that clips updates to the policy gradient to avoid overly large steps [10], while the entropy-maximization term is also popular in actor–critic algorithms, as it enhances their sample efficiency by increasing the diversity of state coverage, especially in continuous control [11].

There is another difference between model-free and model-based approaches. Model-based methods learn (or are provided with) a model of the environment, which can be used for planning, for generating rollouts, or for augmenting the experience. A learned model can provide extra synthetic experience and so is usually more sample efficient, while a model free method tends to be more robust to model inaccuracies, but requires significantly more real-world interaction.

Several of these algorithmic decisions are guided by the problem of exploration-exploitation: A decision maker has to decide between exploring the value of options that may lead to even better returns in the

future, or exploiting the best option known. This is extended in more sophisticated approaches, such as intrinsic-curiosity signals, which are given for visiting novel state, posterior sampling over candidate value functions, and optimism in the face of uncertainty. The trade-off is more severe if rewards are rare and planning is long term, which occurs in many application areas.

4. CURRENT TRENDS IN REINFORCEMENT LEARNING

This section groups contemporary RL research into six partly overlapping themes: deep reinforcement learning as an algorithmic foundation; model-based methods for improving sample efficiency; offline methods that reduce the need for direct access to the environment; multi-agent extensions in which several agents learn simultaneously and must coordinate; the use of human feedback to guide learning, notably for aligning language models; and safety-oriented methods for high-stakes applications. These themes interact: for instance, an offline policy may be pre-trained and then fine-tuned online, or safety constraints may be imposed on a multi-agent system.

A) Deep Reinforcement Learning.

Using deep neural networks as function approximators proved to be a key development, allowing RL to be applied to problems with large number of states and actions that are represented as continuous real numbers instead of discrete values. Using deep neural networks as function approximators also made RL applicable to problems with large continuous real number state and action spaces, rather than tabular spaces. A combination of the convolutional network and experience replay along with a slowly updated target network was able to play many Atari 2600 games from pixels alone and performed as well as or better than human players on many of the games [2]. It sparked many years of algorithmic development, such as reducing the overestimation bias of the common Q-learning algorithm [9], separating the estimation of value functions for the state and the advantage for the action [12] and prioritized replay aimed at focusing updates on the most informative transitions. On the policy-gradient side, there arose the need to use trust-region methods [14] and their simplified variant, proximal policy optimization [10] for continuous control and for large-scale training, due to their simplicity, training stability, and good empirical performance. More recently, deep RL agents have been shown to beat human players in team based realtime strategy games with long time horizons, sparse rewards, partially observable environments and continuous cooperation between agents like Dota 2 [13]. In the recent years, the three key areas of scalability, sample efficiency, and general applicability have been studied across algorithmic classes and the consensus is that the best current benchmark performers compromise in these three facets: exploration, credit assignment, and representation learning [14, 15, 16]. As the number of algorithms has increased, there have been repeated requests for benchmark suites and repeatability, as reported performance of deep RL can be very sensitive to hyperparameters, random seeds, and implementation details which are not always clearly reported.

B. Model Based RL

With the knowledge of an internal model of system dynamics, the model-based RL algorithm can plan or generate synthetic training data with far less interaction with the real environment. Those include locally linear dynamics models and planners like model-predictive control (MPC) or tree search [17] and deep generative models and transformer models of the world. The techniques are particularly useful when real-world experience is too expensive, too slow or too dangerous, such as in robotics where candidate behaviour can be learned in a virtual world model rather than on real-world hardware [18]. Prediction errors accumulate with longer model rollouts, but the sample-efficiency benefits are still large and uncertainty-aware and ensemble dynamics models are becoming increasingly popular which allow to identify if a policy is outside the region of the model where prediction is reliable [19].

C. Offline (Batch) Reinforcement Learning.

Many current RL studies focus on agents that have to learn a policy from a given data set, with no additional communication with the environment during learning. This offline, or batch, setting is important in practice since, in some application areas like healthcare, financial or industrial control, online exploration is either too expensive, too risky, or simply not possible. The main challenge is distributional shift, where the learned policy will have a different behavior from the behavior policy used for data collection, and the value of the learned policy will be measured at states-action pairs that are not well captured by the data, resulting in poor values that the learned policy can exploit. Explicit or implicit constraints that ensure that the learned policy stays within a specific distribution of actions, conservative value estimation that assigns a high cost to out-of-distribution actions, and sequence-modeling methods that formulate trajectory optimization as supervised prediction of returns conditioned on the sequence of actions. A major challenge in this aspect is the balance between conservativeness for robustness and policy improvement for performance relative to original behavior policy.

D. Multi-Agent Reinforcement Learning (MARL), D.2.4.8. Multi-Agent Reinforcement Learning (MARL)

Numerous real-world systems are multi-agent in nature, such as traffic networks, communication systems, multi-robot teams, and economic markets. Multi-agent reinforcement learning (MARL) is a field that focuses on learning agents that operate in a common environment, can cooperate or compete, and are represented by multiple agents. The shared reward has to be distributed across several agents, and the number of agents can be increased by any number, making exploration more extensive, learning and value estimation more difficult [8]. One of the widely used models is centralized training with decentralized execution, where agents can observe information on the global scale (such as the states of all agents, the actions of all agents) in the training phase but can only observe local information (such as

states of local agents, actions of local agents) in the execution phase. Scalable value-decomposition approaches have seen continued progress, while policies that utilize communication to share learned messages have been reported to work in practice, and game-theoretic formulations with provable convergence guarantees are mostly addressed in restricted settings like two-player zero-sum games, the guarantees for MARL with large populations are still largely missing.

E. Reinforcement Learning from Human Feedback and Language-Model Alignment.

RL is one of the latest and most impactful developments for finetuning large language models to human preferences. It starts with training a reward model to predict human preference ratings for pairs of candidate outputs, as done in previous work [20] that demonstrated that one can learn a deep RL policy from comparative human feedback without a hand-designed reward function. Optimizing a language-model policy using PPO maximizes the learned reward and a penalty is applied for going too far from a reference model. Reinforcement learning from human feedback (RLHF) is a crucial step towards making general-purpose language models into instruction-following assistants and is the basis of several popular conversational systems [6]. RLHF pipelines are computationally costly, unstable, and easily exploited by the policy by taking advantage of the weaknesses in the reward (reward hacking), and there has been much follow-up work that has looked for more straightforward alternatives. A recent method, called direct preference optimization (DPO), avoids the need for a separate reward model and an on-policy RL loop by reinterpreting the RLHF objective as a closed-form supervised loss directly over preference pairs [21]. Some more recent works on RL for large language models [22, 23] elaborate on these variations with automated feedback instead of human feedback, multimodal and multilingual settings, and rule-based rewards for tasks where reasoning and coding can be automatically verified and computed. RL and NLP have brought together a large number of people to the RL field and introduced new challenges in reward design, human supervision on a large scale, and aligning rewards with human goals.

F. Safe and Robust Reinforcement Learning

As RL systems move from simulated benchmarks to real-world, safety-critical deployment, ensuring safe behavior during both training and deployment has become a major research challenge. Safe RL is typically approached in two ways: constraint-based optimization, in which the agent chooses actions that achieve high expected return while keeping expected cumulative cost within a bound, and shielding or filtering, in which an additional safety layer, such as a human overseer, screens or corrects unsafe actions proposed by the learned policy before they are executed. A closely related problem is robustness: policies learned in simulation or from a particular data distribution may lose performance when deployed in the real world under distributional shift or adversarial perturbations, which has motivated work on uncertainty quantification, adversarial training, and domain randomization. Because deep neural



network policies are difficult to analyze formally, certifying RL systems with formal safety guarantees remains a significant obstacle in fields such as aviation, medicine, and autonomous transportation [7].

5. APPLICATION DOMAINS

A. Systems and Operations Research

The field of robotics and autonomous-system control is natural and challenging because of the sequential nature of the tasks, the large state space, the large action space, and the presence of sensor and actuator noise. RL has been successfully used for locomotion, dexterous manipulation, and for the control of legged robots; in most cases, the agent is trained in simulation and then transferred to physical robots, dubbed sim-to-real transfer [18]. It has been reported that training policies with a technique called "domain randomization" that systematically changes the simulator parameters during training, can lead to more robust policies in the real world.

B. Autonomous Driving

RL is being explored to design behaviour-planning policies for autonomous vehicles, intersection navigation policies, and end-to-end controllers that translate sensory data into steering and throttle actions. While these RL-based driving policies have demonstrated significant promise, they present significant safety and verification challenges such as simulating catastrophic failure modes exhaustively and ensuring safe operation of the policies in parallel with human drivers [25].

C. Healthcare

In medicine RL has been used in dynamic treatment regimens [24] where the time dependence of the decisions and the time lag between treatment and results are well suited to the RL paradigm, as well as with personalized dosing and the development of clinical decision support systems [25]. Most of the time, the data that is used to train the RL method used here is retrospective data from electronic health records (EHR), which is why distributional-shift difficulties also occur here, along with the challenge of defining appropriate reward functions.

D. Communication Networks and the Internet of Things

RL has also been used in communication networks, where sequential decision-making under uncertainty naturally occurs in adaptive routing, adaptive resource allocation, congestion control and network slicing. RL can be applied to learn adaptive policies that adapt to traffic changes without using static heuristics in routing and traffic engineering; also, it can be implemented for task offloading and energy-aware task scheduling in resource-limited and heterogeneous systems in Internet of Things (IoT) and edge-computing environments. With the advent of 6G networks, the capability of integrating sensing,

communication, and intelligent control in a tight link has brought back the interest in adaptive decision-making methods such as RL for 6G network management.

E. Games and Simulated Environments.

It is likely that games will continue to be a prominent area of RL research in the future because the reward function is well-defined, the cost and degree of parallelism of simulation is low, and it is easy to compare the performance of RL algorithms against human performance. RL algorithms have also been successful in complex partially-observable games, where players must plan for the long-term and coordinate with other agents in the team [13] and more recently in board games like Go [3]. These games will continue to provide playgrounds for new algorithms, including multi-agent learning and multi-agent exploration.

6. CHALLENGES AND OPEN PROBLEMS

Even though significant strides have been made, there are still a number of obstacles to the regular use of RL. Sample efficiency is crucial, but most deep RL algorithms need a much larger number of interaction samples than can be obtained in real world application, which is a motivation for model-based RL [17], transfer learning and offline RL [26]. Reward specification is another challenge: When the reward function is poorly designed, the agent may maximize the reward that was designed but not the reward that it is trying to maximize. In particular, this problem is striking in language models trained with learned rewards [6, 21]. It also is not as flexible as human learning because generalization across tasks and environments is generally weak, and is considered one of the more rudimentary structural limitations of current deep RL architectures [27] policies trained in one environment do not necessarily transfer to another similar, but different, environment. In the case of few or distant rewards, exploration is still hard. Last but not least, policies are still not easily interpretable enough to enable formal verification required in safety-critical application areas [7].

7. FUTURE DIRECTIONS

There are several directions that seem to be likely to influence the future of RL research. Second, with increased integration with large-scale language, vision, and multimodal foundation models [22] and [23], it is possible that the RL agents will start using such pre-trained representations, relieving some of the sample-efficiency limitations discussed in Section V. Second, hybrid approaches that integrate model based and model free learning and uncertainty-aware planning are likely to gain the attention of future researchers because they can leverage the ability of the learned model to achieve sample efficiency with the robustness of the model free approach to policy optimization [19]. Third, offline-to-online RL, where a preliminary policy is first learned offline before being fine-tuned online, can be a safer learning paradigm than fully offline than fully online and could be more common in industry like healthcare or

industrial control [5]. Lastly, before RL can be widely used in safety-critical, heavily-regulated industries like aviation, healthcare, and industrial automation, additional progress is required in safe and robust RL methods, such as using formal verification to validate neural policies and developing shielding architectures that provide formal assurances. The scalability of human oversight—how to maintain helpfulness and calibration of feedback as trained systems grow in capacity to produce higher-quality results than individual evaluators—will be another major research challenge as feedback reaches larger and larger audiences, and as RLHF and its variants evolve [23]. Finally, learning from expert demonstrations is also likely to continue being a practical approach to minimizing the cost of exploration, especially when expert data is scarce, and can be used in combination with RL fine-tuning [28]. Last, in the field of communication networks and particularly the adaptive routing and resource management, multi-agent and model-based systems are expected to continue to be relevant due to the distributed and dynamic nature of 6G and IoT-oriented systems [15].

8. CONCLUSION

Reinforcement learning has grown from a comparatively limited field of machine learning that is primarily used to solve small tabular control problem to a growing research field covering language-model alignment, communication networks, robotics, medical applications, and deep learning. The most prominent recent trends and their application areas are discussed in this paper, such as deep RL, model-based RL, offline RL, multi-agent RL, RLHF-based alignment and safe RL, as well as the obstacles obstructing broader use of the most promising directions. With the progress made in algorithm design and computational power, the issues of sample efficiency, reward engineering, generalization, and safety are still open problems that will continue to drive the research agenda for the next few years. Synergies between reinforcement learning and other fields such as large-scale representation learning and formal verification are likely to remain an important factor affecting the field.

9. REFERENCES

- [1] R. S. Sutton and A. G. Barto, “Reinforcement Learning: An Introduction,” 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [2] V. Mnih et al., “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, pp. 529–533, Feb. 2015.
- [3] D. Silver et al., “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, pp. 484–489, Jan. 2016.
- [4] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [5] S. Levine, A. Kumar, G. Tucker, and J. Fu, “Offline reinforcement learning: Tutorial, review, and perspectives on open problems,” arXiv:2005.01643, 2020.



- [6] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2022.
- [7] J. García and F. Fernández, “A comprehensive survey on safe reinforcement learning,” J. Mach. Learn. Res., vol. 16, no. 1, pp. 1437–1480, 2015.
- [8] S. Gronauer and K. Diepold, “Multi-agent deep reinforcement learning: A survey,” Artif. Intell. Rev., vol. 55, no. 2, pp. 895–943, Feb. 2022.
- [9] Y. Li, “Deep reinforcement learning: An overview,” arXiv:1701.07274, 2017.
- [10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” arXiv:1707.06347, 2017.
- [11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 1861–1870.
- [12] X. Wang et al., “Deep reinforcement learning: A survey,” IEEE Trans. Neural Netw. Learn. Syst., vol. 35, no. 4, pp. 5064–5078, Apr. 2024.
- [13] OpenAI et al., “Dota 2 with large scale deep reinforcement learning,” arXiv:1912.06680, 2019.
- [14] A. K. Shakya, G. Pillai, and S. Chakrabarty, “Reinforcement learning algorithms: A brief survey,” Expert Syst. Appl., vol. 231, art. no. 120495, 2023.
- [15] B. Kommey, O. J. Isaac, E. Tamakloe, and D. Opoku, “A reinforcement learning review: Past acts, present facts and future prospects,” IT J. Res. Develop., vol. 8, no. 2, pp. 120–139, Mar. 2024.
- [16] M. Ghasemi, A. H. Moosavi, and D. Ebrahimi, “A comprehensive survey of reinforcement learning: From algorithms to practical challenges,” arXiv:2411.18892, 2024.
- [17] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, “Model-based reinforcement learning: A survey,” Found. Trends Mach. Learn., vol. 16, no. 1, pp. 1–118, 2023.
- [18] A. S. Polydoros and L. Nalpantidis, “Survey of model-based reinforcement learning: Applications on robotics,” J. Intell. Robot. Syst., vol. 86, no. 2, pp. 153–173, 2017.
- [19] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, “A survey on model-based reinforcement learning,” Sci. China Inf. Sci., vol. 67, no. 2, art. no. 121101, 2024.
- [20] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, “Deep reinforcement learning from human preferences,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4299–4307.
- [21] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, “Direct preference optimization: Your language model is secretly a reward model,” in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2023.
- [22] S. S. Srivastava and V. Aggarwal, “A technical survey of reinforcement learning techniques for large language models,” arXiv:2507.04136, 2025.
- [23] R. Sharma, M. Mehta, and S. T. Raina, “RLHF: A comprehensive survey for cultural, multimodal and low latency alignment methods,” arXiv:2511.03939, 2025.



- [24] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artif. Intell. Med.*, vol. 109, art. no. 101964, 2020.
- [25] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.
- [26] H.-n. Wang et al., "Deep reinforcement learning: A survey," *Front. Inf. Technol. Electron. Eng.*, vol. 21, no. 12, pp. 1726–1744, 2020.
- [27] A. Mohan, A. Zhang, and M. Lindauer, "Structure in deep reinforcement learning: A survey and open problems," *J. Artif. Intell. Res.*, vol. 79, pp. 1167–1236, 2024.
- [28] J. Ramírez, W. Yu, and A. Perrusquía, "Model-free reinforcement learning from expert demonstrations: A survey," *Artif. Intell. Rev.*, vol. 55, no. 4, pp. 3213–3241, 2022.