Applications of Reinforcement Learning and its Extension to Tactical Simulation Technologies

Arif Furkan Mendi [1], Dilara Doğan [1], Tolga Erol [1], Turan Topaloğlu [1], M. Esat Kalfaoğlu [2], Hüseyin Oktay Altun [2]
[1] Simulation, Autonomous and Platform Management Technologies, HAVELSAN, Ankara, Turkey.
[2] AutoDidactic Technologies, Konya, Turkey.

Email: afmendi@havelsan.com.tr; ddogan@havelsan.com.tr; terol@havelsan.com.tr; ttopaloglu@havelsan.com.tr; esat.kalfaoglu@autodidactic.com.tr; oktay.altun@autodidactic.com.tr

***Abstract* - Reinforcement Learning (RL) is a branch of machine learning that is used in many areas, from robotics to natural language processing, from game technologies to medical fields and finance. It is widely used in systems containing large data, where instantaneous data flow is intense and where data tagging is arduous or impossible. RL is a preferred approach for exploring new strategies or combinations due to its convenient nature to real life control problems where sequential decision-making is crucial. In the early stages of its development, RL was mainly prevalent in game technologies. However, recently applications of RL have diversified and extended to a plethora of new fields. As HAVELSAN, the leading Turkish defence industry software and simulation company, we investing this technique to take the eye-catching advantages. In this study, we elaborate on the fundamentals of reinforcement learning technology with an emphasis of its novel applications and future projections in the light of existing research findings. We then cover the RL specifically from simulation technologies perspective and introduce the FIVE-ML project of HAVELSAN which aims a transition from rule-based behaviour modelling to learning-based smart behaviour modelling in order to provide more effective and more dynamic pilot training environment.**

Keywords - Simulation technologies, machine learning, deep reinforcement learning, smart behaviour modelling

I. INTRODUCTION

Reinforcement learning (RL) emerges as a new but effective method in many fields from robotics to production systems, from health to natural language processing, from game technologies to finance and military projects. RL is indispensable when the environment we want to learn provides delayed consequences for the actions we take, and needs optimal sequential decision making. RL follows a different approach in terms of defining and solving problems compared to the supervised and unsupervised learning methods in the fields of computer science especially utilizing Bellman equations in the formulation. In addition to the different approaches it follows, there are also studies that are used as hybrid with RL methods, as well as supervised and unsupervised learning methods. Furthermore, RL studies supported by deep learning are also widely used today. According to the Hype Cycle technology reports by Gartner, one of the leading research organizations, it is seen that the machine learning field has been picked up for the last 6 years [1]. However, in Gartner's "Hype Cycle for Data Science and Machine Learning, 2020" report, it was stated that RL and online learning working on real-time data under the umbrella of adaptive machine learning are on the rise. The results of analyses described in detail in the report are summarized with the graphic given in Figure 1 [2].

HAVELSAN, which is a national and international leading system integrator company with advanced technology-based software-intensive original solutions and products in the fields of defence, security and informatics, not only in command control systems, country security and cyber security and information technologies, but also in machine learning, emerging technologies such as artificial intelligence, augmented reality and innovative simulation technologies are also actively studied.

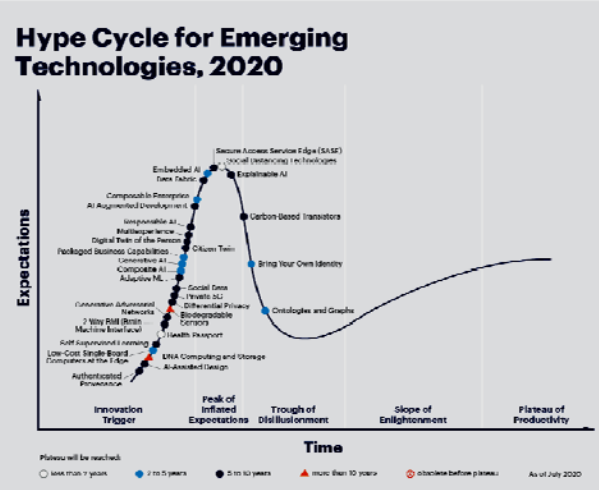

Figure 1. Gartner Hype Cycle for data science and machine learning, 2020 [2]

For this purpose, its simulation, autonomous and platform management technologies department produce high quality civil and military aviation flight simulators, training

supplements, systems for decision support and war gaming in many fields. In addition, studies such as the production of unmanned land vehicles on robotic systems are carried out. With the Forces in Virtual Environment (FIVE) product, one of the simulation software products produced by HAVELSAN, the behaviour of virtual land, air and naval platforms that can perform the defence and offensive tasks required for the tactical training of the simulators of combat platforms can be modelled. Thus, it provides the necessary tactical environments for simulators. With FIVE, simulator users can perform comprehensive tactical and operational training in a safe environment, using all kinds of warfare equipment (e.g., weapons, sensors, countermeasures, electronic warfare, communication and link systems, etc.), without risk and cost effectively. Efforts have been made to make this software, which works autonomous rule-based, based on learning-based artificial intelligence.

The artificial intelligence approach of Forces in Virtual Environment (FIVE-ML) project implemented using RL in the field of simulation technologies with the support of other machine learning strategies such as supervised learning. With the development of the rule-based autonomous behaviour structure of the FIVE software product, developed by HAVELSAN, and the transition to artificial intelligence infrastructure that learns with RL, a cost-effective result has been achieved, which makes a difference to its counterparts in the market. Effective solutions are produced to make many such products better quality and equipped for the needs that may occur with the use of new technologies. In this study, we will explain RL technology and examine the applications of it in various field. We will present our predictions for the future with the perspective that the examined studies and the current situation have given us.

II. RELATED WORKS

There are studies on various subjects examining the studies in the field of reinforcement learning. Robotics, game technologies, health and transportation are among the fields of practice. In a study conducted to control traffic lights and to solve the problem of road congestion at intersections, it was seen that higher success was achieved with the use of multi-agent adaptive reinforcement learning in a simulated environment compared to traditional methods [3]. The reward in the system is defined as minimizing the junction delay. The use of reinforcement learning with different machine learning and deep learning techniques has also become widespread. A pioneer study in this field is the work done to use Deepmind together with Convolutional Neural Network (CNN) and RL [4], where CNN is utilized for feature extraction from an image. Utilizing RL and Recurrent Neural Network (RNN) together is another example to joint utilization of RL and deep learning. t. RNNs are deep learning models that are utilized for sequential learning and temporal modelling. The mathematical background of RL mainly depends on the Markov Decision Process (MDP) where the probability of the next state is only dependent on the current state and the implemented actions. However, this might not be possible in complex environments where past states are also important, resulting Partially Observable Markov Decision Process (POMDP). For example, LSTM is combined with RL to create Deep Recurrent Q-Network (DRQN) to play Atari 2600 games [5]. Additionally, the benefits of this is also shown with the combined utilization of Asynchronous Actor Critic (A3C) [6] with LSTM.

In the study on reinforcement learning difficulties in the field of health, defining the situation and action area, learning and evaluating policies from observational data, and designing reward functions are discussed in terms of challenges [7]. An approach in the field of energy discusses the resolution of energy system-related problems using reinforcement learning are clustered into groups which are related to specific control problems, such as building energy management, dispatch, energy systems in hybrid vehicles, energy markets, grid, and energy devices [8]. The use of AI agents by Deepmind to cool Google Data Centers is a good example [9]. With this study, an energy saving of 40% has been achieved. Centers can now be fully controlled by the AI system without the need for human intervention. In a study published by the Alibaba Group, a multi-agent reinforcement learning approach was applied for the Distributed Coordinated Multi-Agent Bidding (DCMAB) solution, which is used for real-time advertising display offers of Taobao (taobao.com) advertising platform. With the approach applied, the computational complexity between customers who offer advertising and sellers is aimed to reduce [10].

III. WHAT IS REINFORCEMENT LEARNING?

RL is learning what to do to maximize reward and how to match situations to actions. The agent is not told what actions to take, instead, it is asked to discover by trying which actions are the most rewarding actions [11]. RL algorithms are a learning system that interacts with an environment and a feedback loop between past experiences and present [12]. RL working principle is summarized in Figure 2.

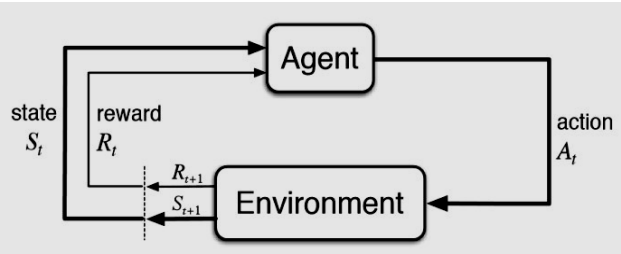

Figure 2. RL work flow principle [11]

In many complex areas, RL is the only viable way to train a program to perform at high levels. For example, when playing games, it is extremely difficult for a person to provide accurate and consistent evaluations of the many

positions that will be needed to train a person with an evaluation function through direct examples. Instead, the program can be told when it won and lost, and it can use this information to learn an evaluation function that gives reasonably accurate estimates of the probability of winning from any position [13]. Alpha Go, developed by Google, where training activities were carried out as mentioned above, helped to hear the success of RL by outperforming the best Go player in the world. Some popular examples of RL algorithms include Q-learning, temporal difference models, and deep RL [14]–[17]. RL performs learning activities with its own experiences, with a focus on reward and punishment, without directly using any labelled data. Initially, it acts to recognize the current situation and conditions and to discover the sources of benefit that it will provide within the system it is in. It acts in a way to optimize the "loss / cost function" in machine learning terminology with the experience it has gained at the end of many experiments. The critical point here is that the more experience and experience gained, the better recognition and understanding brings success with it.

IV. APPLICATION CASES

The usage areas of RL, which can learn very complex behaviours without the need for labelled training data and make short-term decisions while optimizing for a long-term goal, are expanding as time goes on. We will examine some of application areas under the popular topics of machine learning, healthcare and simulation technologies.

A. Popular Machine Learning Topics

1) Natural Language Processing (NLP): Within the scope of natural language processing, data science and machine learning techniques are actively used as technological infrastructure.

One of the important problems addressed in the field of NLP is text summarization. A combination of supervised and RL was used for text summarization in this work [18]. The purpose of the study is to solve the problem encountered in summarizing when using RNN-based encoder-decoder models in longer documents. In this context, it proposes a new intra-attentive neural network that participates in input and continuously produces separate output. It consists of a combination of standard supervised word prediction and RL as a training method.

Another crucial utilization of RL in NLP tasks is for the utilization of non-differentiable NLP metrics such as BLEU and CIDEr during the training of NLP related tasks [19], [20]. Before RL, cross entropy loss is utilized in supervised learning fashion, but this loss does not take into account the coherent relations between the predicted words, reducing the efficiency of the predictions.

Another NLP problem is machine translation is defined as the problem of automatic translation of a given text into a different language by the machine. The Google Translate service provided by Google is a very successful example in this field. An important study in the field of machine translation, an approach based on RL is proposed for simultaneous machine translation [21]. The difference of this study is that it has the ability to learn when to trust predicted words and uses RL to determine when to wait for further input.

Chatbots are simply expressed as computer software that can talk or correspond with people in a digital environment like humans. Chatbots take place in messaging platforms such as WhatsApp, Facebook Messenger, voice assistants such as Google Assistant and Siri, as well as on the relevant institution's own website or mobile application. A study by deep RL for use in dialogue building describes how it integrates these goals by applying deep RL to model future reward in chatbot dialogue [22]. RL model was used in order to model the future direction of a dialogue and to produce coherent and interesting dialogues.

2) Computer Vision & Robotics

Machine learning techniques are actively used in the fields of image processing and computer vision. RL, which is the approach of maximizing the reward with the experiences gained from past experiences in the subjects of perception and analysis of the external world in robots, and extraction of image attributes, has become widespread.

Facebook has created an open-source RL platform called Horizon that can be used for engineering applications [23]. The platform uses RL to optimize large-scale production systems. Horizon has been used by Facebook to customize recommendations, provide more meaningful notifications to users, and optimize video streaming quality.

Robots are more widely used today for increasingly complex purposes such as complex assembly, picking and packaging, last mile delivery, environmental monitoring, search and rescue, and assisted surgery. AWS SageMaker is a service that enables model building, training and testing activities in the field of machine learning, also in a distributed state [24].

B. Medicine & Healthcare

The importance of healthcare has become even more understandable, especially during the COVID-19 pandemic period we are in. In addition, the technologies used in the field of health have increased in importance. Technology is used in many areas such as vaccine-laboratory studies, patient data and follow-up, use of various devices for diagnosis and treatment. Studies have been conducted in the field of health with RL, which offers an innovative machine learning approach.

1) Surgery: A study continues to grow suddenly, providing insight into RL difficulties in the medical field

[25]. The study focuses on the difficulties in formulating the reward function, which defines the ultimate goal and determination of patient states from electronic health records, as well as the lack of resources to simulate the potential benefits of actions proposed in response to changing physiological conditions during and after surgery. This study, which deals with RL in the field of medicine, emphasized that the development and verification of personalized RL models in surgery will contribute to the improvement of care by helping patients and clinicians to make decisions that are more accurate.

2) Early Diagnosis: KenSci uses RL to predict diseases and treatments to help healthcare professionals and patients intervene at earlier stages [26]. It also allows you to predict various health threats that can affect the population by disease progression, detecting patterns, and creating precarious signs.

3) Pharmacy Personalization Treatment: Optimization of anaemia management in patients with chronic kidney failure is presented. The goal is to customize the treatment, i.e., erythropoietin dosages, to stabilize patients in the targeted hemoglobin (Hb) range. The results show that the use of RL increases the proportion of patients in the desired Hb range. Thus, the quality of life of patients increases and the health system reduces costs in anaemia management. This is a relevant problem in Nephrology, in which we focus on obtaining the optimal erythropoietin (EPO) dosages that should be administered for an adequate long term anaemia management [27].

4) Chemistry: The model created by using the deep RL technique for the optimization of chemical reactions in a study conducted in the field of chemistry, proceeds by choosing new conditions based on the experiences obtained by repeatedly recording the results of chemical reactions [28]. RL used with LSTM to model policy, Markov decision process (MDP) was used for chemical reaction optimization. It demonstrated a more successful approach than the best-known black box algorithm in 71% fewer steps in real reactions.

C. Simulation Technologies

Simulation is an attempt to model a real-life or hypothetical situation on a computer so that it can be studied to see how the system works. The simulation system consists of a model of systems or processes that contain defined relationships between their objects. In simulation technologies, which are defined as modelling a real entity or system in simulation technologies, it is possible to maximize the benefit by learning from experience, that is, the use of RL technology. With this understanding, various studies have been carried out using RL technology.

DARPA, an agency of the US Department of Defence responsible for developing new technologies for use by the US military, has published the Air Combat Evolution (ACE) program. ACE is an important proof that autonomous warfare technology can be relied upon by defeating the AI champion fighter pilot in a human-machine cooperative dogfight as the first challenge scenario [29].

In another study, in order for simulated warplanes to fight each other in a virtual war, the tactical decisions of the aircraft were made with RL [30]. Intelligent virtual assets and their behavioural factors have been created through information representation and processing by tactical information bases, creating air combat experience and models. In this way, routes similar to the flight routes preferred by real pilots during the war were drawn and these routes were monitored in a simulation environment on smart assets.

In a different study that allowed pilots to be trained in specific combat tactics in a combat simulation system, a structure was created that allows pilots to create strategies in different situations and then establish the relationship between machine learning input and output [31]. In this study, using a method similar to actor-critic learning, one of the deep RL methods, it trains a radiant neural network model by learning the simulation data respectively, and then makes real-time predictions.

In real scenarios and systems, there is an approach to act as a team rather than as a single entity. A decision-making mechanism has been proposed for missile target assignment using the particle swarm optimization algorithm technique that can be used in such scenarios [32]. With an approach similar to the policy optimization in RL, a system that learns the speed of the aircraft and can detect multiple targets faster, has been designed.

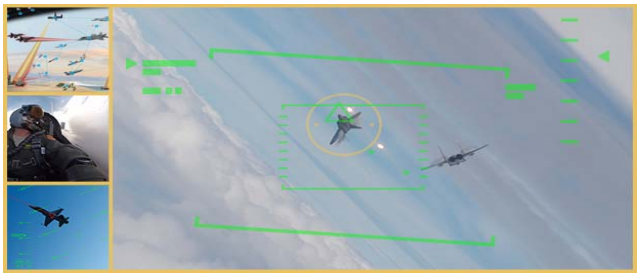

Figure 3. A snapshot from DARPA Dogfight [23]

In the literature, RL is also considered specifically for increasing the effectiveness of pilot training [33]. In this study, live, virtual and constructive (LVC) paradigm is addressed. It is concluded that multi-agent multi-objective RL can create pilot behaviour models can cooperate with each other and prioritize the conflicting objectives in air combat scenarios. However, it is also denoted that more improvement is a must to develop fully professional behaviour models for air combat scenarios. Moreover, there are some works that focuses on the intelligent manoeuvre

model of unmanned aerial vehicles (UAV) which are crucial for future autonomous aerial combats [34], [35]. Manoeuvre model is a continuous domain problem and it is seen that DDPG [14] is chosen for this problem in these studies.

FIVE product, which is developed by HAVELSAN, is used in various simulation projects in Turkish defence industry. With this product, the behaviour of virtual land, air and naval platforms that can perform the defence and offensive tasks required for the tactical training of the simulators of combat platforms can be modelled, thus providing the necessary tactical environments for simulators. With the FIVE, simulator users; can perform comprehensive tactical and operational training in a safe environment, using all kinds of warfare equipment (weapons, sensors, countermeasures, electronic warfare, communication and link systems, etc.) without risk and cost. Similarly, in the development process of combat platforms; analysis and verification activities, which are expected to be very costly and risky in analysis, design and testing stages, can be carried out cost-effectively and safely with FIVE. Currently, there are 273 platforms and 400 subsystems defined on FIVE. There will be an increase in the number of systems and subsystems over time. The actions taken on the players in the product infrastructure are carried out with rules. For the correct use of the functions used in defining these rules, it requires field knowledge as well as programming. For this reason, the need to work closely with field experts has emerged.

Figure 4. Virtual tactical environment representation [36]

The transition to a new approach based on RL has been realized in order to minimize this need, to transform deterministic models into a learning and developing state with the use of artificial intelligence techniques, and to increase the training quality of pilots who receive training using the product. HAVELSAN, which produces D-level simulators according to world standards in the field of simulation, has achieved one of the first important applications of RL in the field of simulation with the tactical environment simulator project (FIVE-ML), where it differentiates its counterparts by using RL technology. FIVE-ML, various scenarios involving the tactical movements of main air-to-air platforms and air to land platforms under various conditions have been studied. By defining and using different combinations of these scenarios and using reward-punishment functions, trainings were made with RL and training activities were carried out for the players.

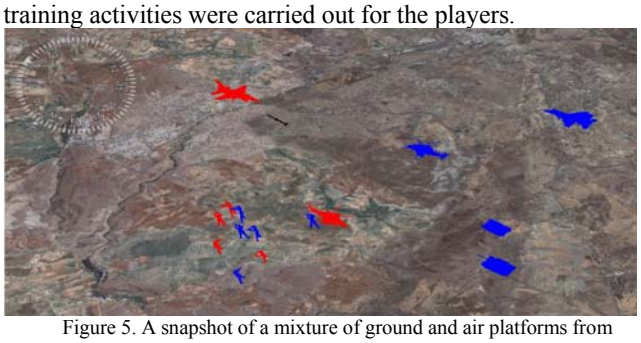

Figure 5. A snapshot of a mixture of ground and air platforms from FIVE

Within the scope of this study, air to air and air to ground engagement scenarios including up to six virtual entities are studied. The scenarios cover three friend and three enemy entities. RL-based behaviour models control two of the three friend entities. The other entities are controlled by currently existing HAVELSAN rule-based behaviour rules. It is seen that for the same starting position scenarios, RL based entities win against the HAVELSAN rule-based entities. During these experiments, it is also observed that utilizing supervised learning as an initial point to RL reduces the training time significantly and creates more realistic behaviour models. The supervised learning models are trained with currently existing rule-based HAVELSAN rules. One of the results of the RL training is shown in Figure 6. FIVE-ML will possibly open the way to decision support systems that can be used in the real operation environment. It is also expected to be used actively as decision support systems, which have an important strategic and military place.

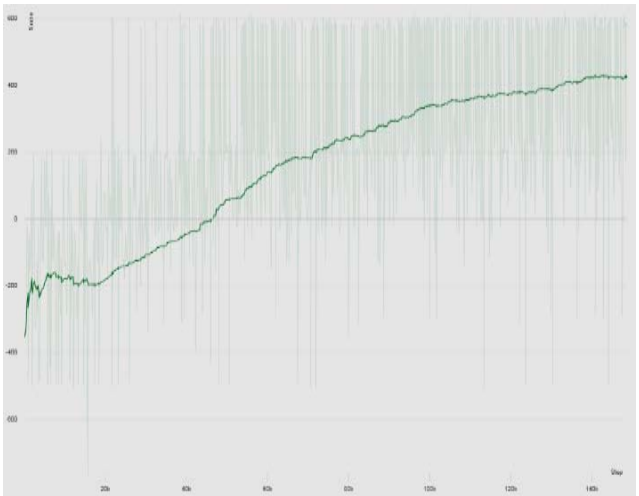

Figure 6. The smoothed score graph of RL training on air-to-ground scenario in FIVE software. Score is the total reward obtained at the end of the episode and there are various rewards such as own hit punishment, enemy hit reward, friend hit punishment, and firing missile punishment etc.

It can be emphasized that the FIVE-ML project, which was realized with the knowledge and experience we have

gained as HAVELSAN, and the use of RL in the field of simulation will become important points to go for simulation technologies. Considering that, the importance of defence technologies and decision support systems that enable strategic decisions to be made rapidly, the FIVE-ML project will take its place in the literature as a state-of-the-art study with the use of RL technique in the field of simulation. FIVE-ML is an important example and milestone of RL usage since it can be predicted that the success of this work triggers a wide range of projects/products in the simulation technologies field.

V. CONCLUSION

RL is an innovative and developing field, so many studies have been carried out in various a wide range. When we examine the studies in which the RL technique is used in areas such as popular machine learning, health, and simulation technologies, we see that the technology addresses a wide range of uses. Since simulation technology is our main activity area and consists of a wide range of disciplines we wanted to put the emphasis on the studies in simulation technologies, which is one of the most important areas of RL. HAVELSAN, which has become a world brand in simulation technologies, has software-intensive original solutions and products, acts with the vision of being a national and international leading system integrator in its field, and follows an approach that will ensure maximum benefit from developing technologies with the use of emerging technologies in products. In this context, with the FIVE-ML project, a pioneering study has been made for the use of RL in the field of simulation by combining simulation technologies with artificial intelligence. This study, which is an example in the field of simulation, has created a technical infrastructure that can be used in various simulation products and it is predicted that it will be actively used as a decision support system in real operation environments in the near future. As HAVELSAN, with our experience and expertise in simulation, we can easily foresee that RL technology is critical for simulation technologies and will gain even more value in the coming periods. In addition, we anticipate that the work done will diversify and the use of technology will become widespread as the level of maturity of technology increases. In the future, uses for RL in areas such as industry, education, security and smart cities can be examined. As a future work, the study will be expanded to other scenario types such as air-to-air defence, air to ground SEAD, ground to air defence. Moreover, the complexity of the scenarios is planned to increase by increasing the number of virtual entities and RL controlled virtual entities. Moreover, comprehensive trainings will be implemented in order to provide a better robustness against the changes in the type and class of the enemy entities, the changes in the location, heading and altitude of all entities.

REFERENCES

[1] J. Kotecki, "Deep Learning's 'Permanent Peak' On Gartner's Hype Cycle | by James Kotecki | Machine Learning in Practice | Medium," 17-Aug-2018. [Online]. Available: https://medium.com/machine-learning-in-practice/deep-learnings-permanent-peak-on-gartner-s-hype-cycle-96157a1736e. [Accessed: 14-Mar-2021].

[2] S. Vashisth, A. Linden, J. Hare, and P. den Hamer, "Hype Cycle for Data Science and Machine Learning, 2020," 28-Jul-2020. [Online]. Available: https://www.gartner.com/en/documents/3988118/hype-cycle-for-data-science-and-machine-learning-2020. [Accessed: 14-Mar-2021].

[3] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, "Reinforcement learning-based multi-agent system for network traffic signal control," 2009.

[4] V. Mnih *et al.*, "Playing Atari with Deep Reinforcement Learning," 2013.

[5] M. Hausknecht and P. Stone, "Deep Recurrent Q-Learning for Partially Observable MDPs," 2015.

[6] V. Mnih *et al.*, "Asynchronous Methods for Deep Reinforcement Learning," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 4, pp. 2850–2869, Feb. 2016.

[7] E. Riachi, M. Mamdani, M. Fralick, and F. Rudzicz, "Challenges for Reinforcement Learning in Healthcare," Mar. 2021.

[8] A. T. D. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renew. Sustain. Energy Rev.*, vol. 137, p. 110618, Mar. 2021.

[9] DeepMind, "Safety-first AI for autonomous data centre cooling and industrial control | DeepMind," 17-Aug-2018. [Online]. Available: https://deepmind.com/blog/article/safety-first-ai-autonomous-data-centre-cooling-and-industrial-control. [Accessed: 16-Mar-2021].

[10] J. Jin, C. Song, H. Li, K. Gai, J. Wang, and W. Zhang, "Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising," 2018.

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: A Bradford Book, 2018.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, 2016.

[13] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. USA: Prentice Hall Press, 2009.

[14] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[15] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," *NeurIPS*, pp. 1–9, Jan. 2018.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal Policy Optimization Algorithms," Jul. 2017.

[17] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust Region Policy Optimization," *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 3, pp. 1889–1897, Feb. 2015.

[18] R. Paulus, C. Xiong, and R. Socher, "A DEEP REINFORCED MODEL FOR ABSTRACTIVE SUMMARIZATION," 2017.

[19] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical Sequence Training for Image Captioning," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 1179–1195, Dec. 2016.

[20] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-Memory Transformer for Image Captioning," Dec. 2019.

[21] A. Grissom II, H. He, J. Boyd-Graber, J. Morgan, and H. Daumé III, "Don't Until the Final Verb Wait: Reinforcement Learning for Simultaneous Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1342–1352.

[22] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep Reinforcement Learning for Dialogue Generation," 2016.

[23] J. Gauci, E. Conti, and K. Virochsiri, "Horizon: An open-source reinforcement learning platform - Facebook Engineering," 01-Nov-2018. [Online]. Available: https://engineering.fb.com/2018/11/01/ml-

applications/horizon/. [Accessed: 16-Mar-2021].

[24] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-End Training of Deep Visuomotor Policies," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1334–1373, Jan. 2016.

[25] A. Chung *et al.*, "Introducing Amazon SageMaker Reinforcement Learning Components for open-source Kubeflow pipelines | AWS Machine Learning Blog," 03-Mar-2021. [Online]. Available: https://aws.amazon.com/tr/blogs/machine-learning/introducing-amazon-sagemaker-reinforcement-learning-components-for-open-source-kubeflow-pipelines/. [Accessed: 16-Mar-2021].

[26] S. Datta *et al.*, "Reinforcement learning in surgery," *Surg. (United States)*, 2021.

[27] KenSci, "KenSci | AI Platform for Digital Health." [Online]. Available: https://www.kensci.com/. [Accessed: 16-Mar-2021].

[28] Z. Zhou, X. Li, and R. N. Zare, "Optimizing Chemical Reactions with Deep Reinforcement Learning," 2017.

[29] DARPA, "Training AI to Win a Dogfight," 05-Aug-2019. [Online]. Available: https://www.darpa.mil/news-events/2019-05-08. [Accessed: 23-Mar-2021].

[30] X. Lei, A. Huang, T. Zhao, Y. Su, and C. Ren, "A New Machine Learning Framework for Air Combat Intelligent Virtual Opponent," in *Journal of Physics: Conference Series*, 2018, vol. 1069, no. 1.

[31] M. Yaofei, G. Guanghong, and P. Xiaoyuan, "Cognition behavior model for air combat based on reinforcement learning--《Journal of Beijing University of Aeronautics and Astronautics》2010年04期," *J. Beijing Univ. Aeronaut. Astronatics*, vol. 36, no. 4, pp. 379–383, 2010.

[32] H. GUO, H. XU, X. GU, and D. LIU, "Air Combat Decision-Making for Cooperative Multiple Target Attack Based on Improved Particle Swarm Algorithm," *Fire Control Command Control*, 2011.

[33] J. Källström and F. Heintz, "Multi-Agent Multi-Objective Deep Reinforcement Learning for Efficient and Effective Pilot Training," in *Proceedings of the 10th Aerospace Technology Congress, October 8-9, 2019, Stockholm, Sweden*, 2019, vol. 162, pp. 101–111.

[34] W. Kong, D. Zhou, Z. Yang, Y. Zhao, and K. Zhang, "UAV Autonomous Aerial Combat Maneuver Strategy Generation with Observation Error Based on State-Adversarial Deep Deterministic Policy Gradient and Inverse Reinforcement Learning," *Electronics*, vol. 9, no. 7, p. 1121, Jul. 2020.

[35] Q. Yang, Y. Zhu, J. Zhang, S. Qiao, and J. Liu, "UAV Air Combat Autonomous Maneuver Decision Based on DDPG Algorithm," in *IEEE International Conference on Control and Automation, ICCA*, 2019, vol. 2019-July, pp. 37–42.

[36] Modern Entegrated Warfare, "WHITEPAPER: Exclusive Download: LVC-Enabled Testbed for Autonomous System Testing - Modern Military Training," 14-Dec-2016. [Online]. Available: http://modernmilitarytraining.com/blended-training/lvc-enabled-testbed-autonomous-system-testing/. [Accessed: 23-Mar-2021].