An Invitation to Deep Reinforcement Learning

Other titles in Foundations and Trends® in Optimization

Constrained Reinforcement Learning with Average Reward Objective: Model-Based and Model-Free Algorithms
Vaneet Aggarwal, Washim Uddin Mondal and Qinbo Bai
ISBN: 978-1-63828-396-6

Stochastic Optimization Methods for Policy Evaluation in Reinforcement Learning
Yi Zhou and Shaocong Ma
ISBN: 978-1-63828-370-6

Numerical Methods for Convex Multistage Stochastic Optimization
Guanghui Lan and Alexander Shapiro
ISBN: 978-1-63828-350-8

A Tutorial on Hadamard Semidifferentials
Kenneth Lange
ISBN: 978-1-63828-348-5

Massively Parallel Computation: Algorithms and Applications
Sungjin Im, Ravi Kumar, Silvio Lattanzi, Benjamin Moseley and Sergei Vassilvitskii
ISBN: 978-1-63828-216-7

Information Relaxations and Duality in Stochastic Dynamic Programs: A Review and Tutorial
David B. Brown and James E. Smith
ISBN: 978-1-68083-962-3

An Invitation to Deep Reinforcement Learning

Bernhard Jaeger
University of Tübingen
bernhard.jaeger@uni-tuebingen.de

Andreas Geiger
University of Tübingen
a.geiger@uni-tuebingen.de

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Optimization

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

B. Jaeger and A. Geiger. *An Invitation to Deep Reinforcement Learning*. Foundations and Trends® in Optimization, vol. 7, no. 1, pp. 1–80, 2024.

ISBN: 978-1-63828-441-3

Foundations and Trends® in Optimization
Volume 7, Issue 1, 2024

Editorial Board

Editorial Scope

Foundations and Trends® in Optimization publishes survey and tutorial articles in the following topics:

- algorithm design, analysis, and implementation (especially, on modern computing platforms
- models and modeling systems, new optimization formulations for practical problems
- applications of optimization in machine learning, statistics, and data analysis, signal and image processing, computational economics and finance, engineering design, scheduling and resource allocation, and other areas

Information for Librarians

Foundations and Trends® in Optimization, 2024, Volume 7, 4 issues. ISSN paper version 2167-3888. ISSN online version 2167-3918. Also available as a combined paper and online subscription.

Contents

An Invitation to Deep Reinforcement Learning

Bernhard Jaeger and Andreas Geiger

University of Tübingen, Tübingen AI Center, Germany;
bernhard.jaeger@uni-tuebingen.de, a.geiger@uni-tuebingen.de

ABSTRACT

Training a deep neural network to maximize a target objective has become the standard recipe for successful machine learning over the last decade. These networks can be optimized with supervised learning if the target objective is differentiable. However, this is not the case for many interesting problems. Common objectives like intersection over union (IoU), and bilingual evaluation understudy (BLEU) scores or rewards cannot be optimized with supervised learning. A common workaround is to define differentiable surrogate losses, leading to suboptimal solutions with respect to the actual objective. Reinforcement learning (RL) has emerged as a promising alternative for optimizing deep neural networks to maximize non-differentiable objectives in recent years. Examples include aligning large language models via human feedback, code generation, object detection or control problems. This makes RL techniques relevant to the larger machine learning audience. The subject is, however, time-intensive to approach due to the large range of methods, as well as the often highly theoretical presentation. This monograph takes an alternative approach that is different from classic RL textbooks. Rather than focusing on tabular

Bernhard Jaeger and Andreas Geiger (2024), "An Invitation to Deep Reinforcement Learning", Foundations and Trends® in Optimization: Vol. 7, No. 1, pp 1–80. DOI: 10.1561/2400000049.

problems, we introduce RL as a generalization of supervised learning, which we first apply to non-differentiable objectives and later to temporal problems. Assuming only basic knowledge of supervised learning, the reader will be able to understand state-of-the-art deep RL algorithms like proximal policy optimization (PPO) after reading this tutorial.

1

Introduction

The field of reinforcement learning (RL) is traditionally viewed as the art of learning by trial and error [152]. RL methods were historically developed to solve sequential decision making tasks. The core idea is to deploy an untrained model in an *environment*. This model is called the *policy* and maps inputs to actions. The policy is then improved by randomly attempting different actions and observing an associated feedback signal, called the *reward*. RL techniques have demonstrated remarkable success when applied to popular games. For example, RL produced world-class policies in the games of Go [137], [145], [146], Chess [137], [146], Shogi [137], [146], Starcraft [162], and Stratego [124], and achieved above human level policies in all Atari games [7], [45], [86] as well as Poker [23], [24], [114]. While these techniques work well for games and simulations, their application to practical real-world problems has proven to be more difficult [43]. This has changed in recent years, where a number of breakthroughs have been achieved by transferring RL policies trained in simulation to the real world [14], [38], [88] or by successfully applying RL to problems that were traditionally considered supervised problems [50], [107], [119]. It has long been known that any supervised learning (SL) problem can be reformulated as an

RL problem [12], [82] by defining rewards that match the loss function. This idea has not been used much in practice because the advantage of RL has been unclear, and RL problems have been considered to be harder to solve. A key advantage of RL over SL is that the optimization objective does not need to be differentiable. To see why this property is important, consider the task of text prediction, at which models like ChatGPT had a lot of success recently. The large language models used in this task are pre-trained using self-supervision [25] on a large corpus of internet text, which allows them to generate realistic and linguistically flawless responses to text prompts.

However, self-supervised models like GPT-3 cannot directly be deployed in products because they are not optimized to predict helpful, honest, and harmless answers [9]. So far, the most successful technique to address this problem is called RL from human feedback (RLHF) [9], [32], [119], [151] in which human annotators rank outputs of the model and the task is to maximize this ranking. The mapping between the models outputs and a human ranking is not differentiable, hence SL cannot optimize this objective, whereas RL techniques can. Recently, RL was also able to claim success in code generation [107] by maximizing execution speed of predicted code, discovering new optimization techniques. Execution speed of code can easily be measured, but not computed in a differentiable way. Derivative-free optimization methods [53], [68] can also optimize non-differentiable objectives but typically do not scale well to deep neural networks. A second advantage RL has over SL is that algorithms can collect their own data which allows them to discover novel solutions [107], [145] that a static human annotated dataset might not contain.

The recent success of RL on real world problems makes it likely that RL techniques will become relevant for the broader machine learning audience. However, the field of RL currently has a large entry barrier, requiring a significant time investment to get started. Seminal work in the field [15], [63], [138], [139] often focuses on rigorous theoretical exposition and assumes that the reader is familiar with prior work. Existing textbooks [52], [152] make little assumptions but are extensive in length. Our aim is to provide readers that are familiar with supervised machine learning an easy entry into the field of deep reinforcement learning to

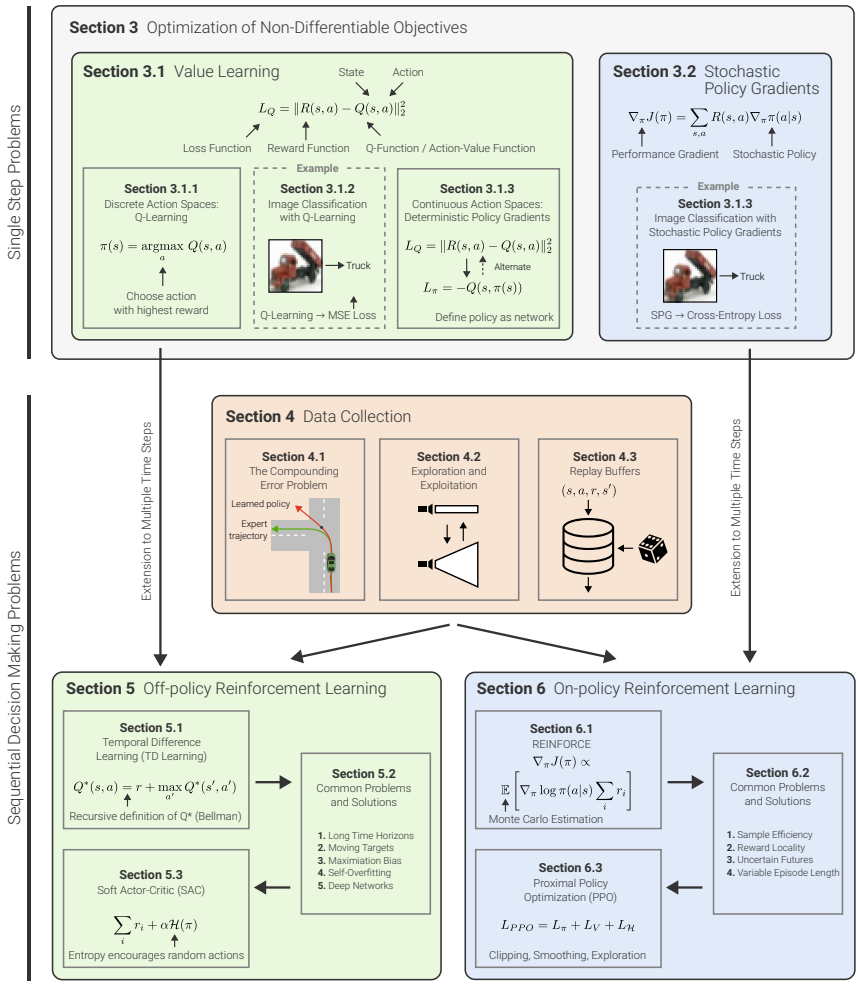

Figure 1.1: An Invitation to Deep Reinforcement Learning. This tutorial is structured as follows: We start by introducing RL techniques through the lens of optimizing non-differentiable metrics for single step problems in Section 3. In particular, we discuss value learning in Section 3.1 and stochastic policy gradients in Section 3.2. For each category of algorithms, we provide a simple example assuming a fixed labeled dataset, thereby connecting RL to SL objectives. This assumption is lifted in Section 4 where we discuss data collection for sequential decision making problems. Next, we extend the techniques from Section 3 to sequential (multi-step) decision making problems. More specifically, we extend value learning to off-policy RL in Section 5 and stochastic policy gradients to on-policy RL in Section 6. For both paradigms, we introduce basic learning algorithms (TD-Learning, REINFORCE), discuss common problems and solutions, and introduce a modern advanced algorithm (SAC, PPO).

facilitate the widespread adoption of these techniques. Towards this goal, we skip the typically rather lengthy introduction via tables and Markov decision processes. Instead, we introduce deep RL through the intuitive lens of optimization. In this monograph, we introduce the reader to all relevant concepts to understand successful modern Deep RL algorithms like proximal policy optimization (PPO) [140] or soft actor-critic (SAC) [63].

Our invitation to RL is structured as follows. After discussing general notation in Section 2, we introduce RL techniques by optimizing non-differentiable metrics in Section 3. We start with the standard supervised setting, e.g., image classification, assuming a fixed labeled dataset. This assumption is lifted in Section 4 where we discuss data collection in sequential decision making problems. In Sections 5 and 6, we will extend the techniques from Section 3 to sequential decision making problems, such as robotic navigation. Figure 1.1 provides a graphical representation of the content.

Appendices

A

Upside Down RL

In the appendices, we briefly introduce some additional topics from the field of RL that are important, but more niche than the ideas covered in the main work.

Upside Down RL [93], [135], [150] is a different but simple concept to bridge the non-differentiable gap between the action and a reward. The idea is to use the reward as conditioning input of the policy:

$$L_\pi := \|a - \pi(s, r)\|_2^2 \tag{A.1}$$

Figure A.1: Upside down RL conditions on the reward.

Additionally, the number of steps in an episode can be added to the input. The policy network is then simply trained with SL, predicting the action that achieved the given reward in this state. In sequential problems, the return can be used for conditioning. Upside Down RL is illustrated in Figure A.1. During inference, the reward is simply set to

the maximum reward to obtain the best action. Upside down RL is a relatively new idea and still part of ongoing research. It has seen the most success when combined with transformers in offline RL settings [30].

B

Model-based Reinforcement Learning

In model-based RL [62], [65]–[67], the non-differentiable environment gap is bridged by learning the environment dynamics explicitly via self-supervised learning. A differentiable model, called the *world model* is optimized to predict the next state and reward, given the current state and action. Compared to model free methods, much richer labels are available because the next state is usually high dimensional. The world model can then for example be used to maximize the return inside the world model directly because it is differentiable. This is illustrated in Figure B.1. Backpropagating through long time horizons can be computationally expensive if the world model has many parameters. A world model can also be used as a learned simulator, which offers a way to generate large amount of samples when environment interaction with the real system is limited. A disadvantage of model-based RL is that the policy can and will exploit inaccuracies in the world model. For example, if the world model incorrectly attributes a lot of reward to an action, the policy trained inside the world model will pick that action even when this action is suboptimal in the real environment. Inaccuracies in the predicted observations can also be problematic if small details in the input are relevant for the downstream task. The

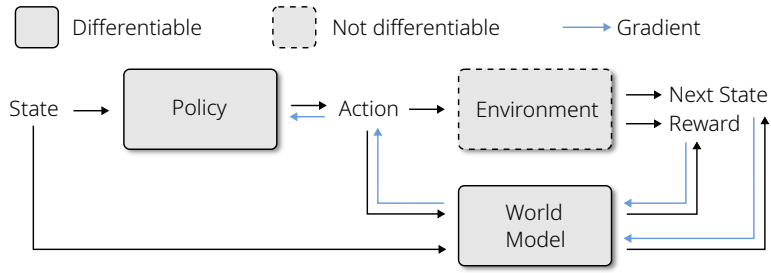

Figure B.1: Model-Based RL learns the environment self-supervised.

world model might not learn small details because they have a low impact on the loss for predicting the next state. Despite the downsides, model-based RL can be useful in settings where the number of available interactions with the real environment is limited.

C

Reinforcement Learning from Human Feedback

Reinforcement Learning from Human Feedback (RLHF) [9], [32], [119], [151] describes the idea of using rankings from human annotators as the target objective to optimize or fine-tune a model. The optimization uses a combination of the standard RL ideas discussed in the main text. RLHF is primarily used to optimize generative models in particular large language models, thus we will focus our discussion on the particular considerations of that task. RLHF has been an integral technique used to turn large language models into useful products like ChatGPT. Similar ideas have also been applied to models that generate images [22], [49], [163].

Large language models (LLM) [25] are trained to predict the probability of the next word, or parts of words called tokens, given prior words in a sentence. This is a self-supervised objective which enables training on internet scale datasets. At inference, these models can be used to generate text by iteratively sampling a word from the predicted distribution. This generates plausible sounding text given an initial text, called the *prompt*. Generating plausible continuations of text can be useful because, for example, the correct answer to a question contains some of the most likely words. However, the correct answer is not the

only plausible continuation of the text. The large scale datasets from the internet that LLMs are trained with also contain lies, offensive speech, manipulative or simply unhelpful text. LLMs trained in a self-supervised way on this data may therefore also generate such responses, and are therefore not safe to deploy into products for end users. One remedy to this problem is *supervised fine-tuning* (SFT) where a labeled dataset with prompts and target texts from a human annotator is collected and trained with SFT has limited effectiveness because creating large labelled datasets with demonstrations is expensive. Additionally, individual human annotators have limited skill sets, for example, they don't know the correct answer to every question for which a correct answer is known and available on the internet. A more scalable approach is to collect a dataset where the pre-trained model generates multiple responses to a given prompt, with its internet scale knowledge base. The human annotators are then tasked to rank these predictions from best to worst. This approach is more scalable because it is easier for humans to verify the correctness of an answer rather than coming up with the correct answer from scratch. However, maximizing human rankings is not a differentiable objective, which is where RL comes to the rescue.

In the version of RLHF proposed by Ouyang *et al.* [119] a reward model is first learned from a dataset containing human rankings. The reward model predicts the ranking given a prompt and an answer sampled from the model. This is a form of value learning, where the reward model can be thought of as a Q-function. Learning this Q-function is very hard because, for example, the Q-function needs to know which of the presented answers is correct, to predict which one the human would prefer. The task is made possible by using a pre-trained LLM as the architecture for the Q-function, with minimal modification to be able to predict rankings. LLMs are probabilistic models, so stochastic policy gradients are used to tune them. In particular, Ouyang *et al.* [119] uses the PPO algorithm discussed in Section 6.3.

RLHF combined with supervised fine-tuning has been found effective enough to deploy LLM chatbots on a large scale. The goal of RLHF is, given a learned distribution, to "unlearn" the parts of the distribution that are considered bad behavior. Current RLHF is far from perfect and an active field of research [129], [174]. Models do not forget all harmful

parts of the distribution and also tend to unlearn useful predictions. This is mitigated by mixing RLHF gradients with gradients from the original self-supervised pre-training Ouyang *et al.* [119]. It is worth noting that with RLHF a generative model is unlikely to learn new behavior as it only reinforces predictions that the generative model has already been capable of generating.

D

Planning

To find the optimal action $a^\star$ for each state s we have primarily considered *policies*, the approach of learning a function π that maps states to actions. There is another approach called planning, which describes *algorithms* that given a model of the environment find the optimal action or improve the actions of a policy.

D.1 Tree Search

A powerful class of algorithms are *search* algorithms, out of which *tree search* is arguably the simplest. Tree search requires a world model that given a state and action can predict the next state and reward. This can be a learned world model, but it does not have to be differentiable, a classic simulator also works. Given a state s, the tree search algorithm computes the next state and stores the reward, for every possible action. In this naive version, the action space has to be discrete. The process of simulating the next time step for every possible action is then repeated for all possible next states from the previous iteration until all branches of this tree have finished in a terminal state. The observed rewards are then used to choose the action from the first iteration based on some criterion, such as highest average return. If the environment has

deterministic state transitions, the action space is discrete and the world model perfect, then this algorithm will find the optimal action $a^\star$. This process will then be repeated for the next state, potentially reusing simulations from prior steps. The difficulty of tree search is that the algorithm will exploit any inaccuracies in the world model, and most importantly it is too slow to run for complex environments. Exhaustively simulating all potential futures is not possible in most cases. In the following section, we will describe a more practical class of search algorithms that use the idea of Monte-Carlo sampling [109] to efficiently choose which futures to evaluate.

D.2 Monte-Carlo Tree Search

The core idea of *Monte-Carlo tree search* (MCTS) [35] is to only explore a part of the full tree by using heuristics and random actions to choose which states and actions to evaluate.

MTCS starts by creating a tree with the current state as the root node and iteratively repeats the following 4 steps until a certain time limit or resource constraint is met.

1. **Selection.** A *tree policy* is used to select a state which still has at least one unexplored action.
2. **Expansion.** An unexplored action in that state is chosen, expanding the tree.
3. **Simulation.** The next action is chosen iteratively by a probabilistic policy until the episode ends.
4. **Backup.** The nodes, up until the node starting the simulation, are updated with the return.

Figure D.1 illustrates these 4 steps. The probabilistic policy, also called the *default policy*, from step 3 can be any probabilistic policy but should be fast to evaluate for the whole process to be efficient, so simple linear layers [145] or just a uniform distribution are used in practice. MCTS may start with an empty tree if the current state is novel. If the current state already was a node in the tree from the previous iteration,

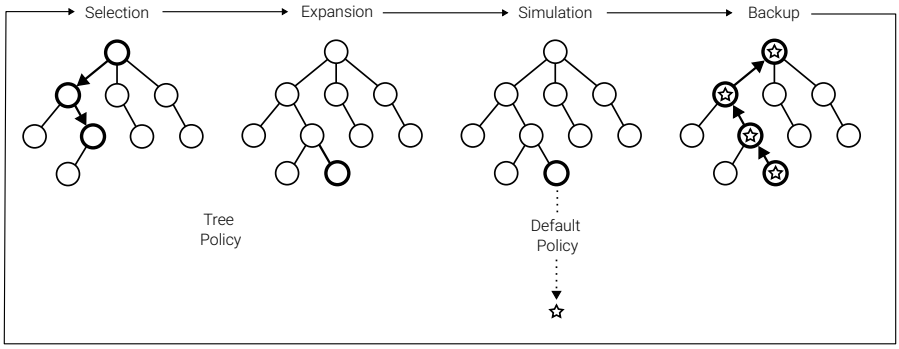

Figure D.1: Monte Carlo Tree Search.

then that node is used as the root node of the new tree and its children are retained. Modern implementations of MCTS combine the idea with policies and value functions trained with RL [145], [146], [148] as well as learned world models [137].

E

Related Work

Many past RL tutorials [56], [69], [100] do not cover deep RL, since this development happened after these works were published. Mousavi *et al.* [115] gives an overview of the deep RL literature, but is shallow in terms of technical details and largely neglects the important topic of policy gradient methods. Levine *et al.* [99] focuses on offline RL, whereas we focus on more mature off-policy and on-policy deep RL techniques. As such, Levine *et al.* [99] is complementary to our work. Vidyasagar [161] focuses on RL theory and does not cover actor-critic methods or modern RL algorithms, whereas we cover modern deep RL.

Surveys on RL [5], [83], [84], [95], [164], [166] typically review the latest research advancements in the field. Our tutorial instead covers ideas that stood the test of time and is as such more suitable as introductory material.

Existing RL Books [52], [102], [106], [152], [154] (as well as lectures [98], [144]) cover a broad range of topics but require a large time commitment to consume. For example, the most widely cited introduction to RL is Sutton *et al.* [152] which is a 526-page-long textbook. It puts a strong focus on theoretical foundations and methods using tabular representations or linear function approximation. For such problems,

much stronger theoretical guarantees can be obtained than for the non-linear function approximation problems that we considered in this work. The textbook also discusses applications of RL in psychology and neuroscience. As such, Sutton *et al.* [152] is complementary to this work, and we recommend it for readers that have an interest in these topics. François-Lavet *et al.* [52] is the closest related manuscript, and can perhaps be seen as a representative of the traditional way to introduce deep RL. It introduces deep RL as techniques for sequential decision making via the theoretical framework of Markov decision processes. On the contrary, we introduce deep RL as a generalization of SL to non-differentiable objectives, which provides an alternative introduction which is more suitable to the larger SL audience. François-Lavet *et al.* [52] covers a broad range of ideas in RL, giving readers a broad overview about many ideas in the field. However, due to covering so many topics it often lacks the necessary depth for the reader to fully understand the presented ideas, requiring the reader to seek out additional material. Instead, our work focuses on depth, by zooming in on the most important ideas covering them in sufficient detail such that the reader can fully understand the ideas and concrete state-of-the-art implementations of them without conducting additional material. As a result, our work is more than 2× shorter and more suitable for readers who want to apply popular RL ideas to domains outside of RL, such as robotics, computer vision or generative AI.

References

[1] J. Achiam, *Simplified PPO-Clip Objective*, URL: https://drive.google.com/file/d/1PDzn9RPvaXjJFZkGeapMHbHGiWWW20Ey/view, 2018.

[2] J. Achiam, *Spinning Up in Deep Reinforcement Learning*, URL: https://spinningup.openai.com, 2018.

[3] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] M. Andrychowicz, A. Raichuk, P. Stanczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, S. Gelly, and O. Bachem, "What matters in on-policy reinforcement learning? A large-scale empirical study," 2020. arXiv: 2006.05990.

[5] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, 2017.

[6] A. Aubret, L. Matignon, and S. Hassas, "A survey on intrinsic motivation in reinforcement learning," 2019. arXiv: 1908.06976.

[7] A. P. Badia, B. Piot, S. Kapturowski, P. Sprechmann, A. Vitvitskyi, Z. D. Guo, and C. Blundell, "Agent57: Outperforming the atari human benchmark," in *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

[8] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, "An actor-critic algorithm for sequence prediction," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017.

[9] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. E. Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022. arXiv: 2204.05862.

[10] L. C. Baird, "Advantage updating," *Technical report wl-tr-93-1146, Wright Patterson AFB OH*, 1993.

[11] B. Bakker, "Reinforcement learning with long short-term memory," in *Advances in Neural Information Processing Systems (NIPS)*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds., 2001.

[12] A. G. Barto and T. G. Dietterich, "Reinforcement learning and its relationship to supervised learning," *Handbook of learning and approximate dynamic programming*, 2004.

[13] S. Beery, G. V. Horn, and P. Perona, "Recognition in terra incognita," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.

[14] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, "Autonomous navigation of stratospheric balloons using reinforcement learning," *Nature*, 2020.

[15] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2017.

[16] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of Artificial Intelligence Research (JAIR)*, 2013.

[17] R. E. Bellman and S. E. Dreyfus, *Applied dynamic programming*. RAND Corporation, 1962.

[18] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2009.

[19] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, "Dota 2 with large scale deep reinforcement learning," 2019. arXiv: 1912.06680.

[20] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] N. Bjorck, C. P. Gomes, and K. Q. Weinberger, "Towards deeper deep reinforcement learning with spectral normalization," *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[22] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine, "Training diffusion models with reinforcement learning," 2023. arXiv: 2305.13301.

[23] N. Brown and T. Sandholm, "Superhuman ai for heads-up no-limit poker: Libratus beats top professionals," *Science*, 2018.

[24] N. Brown and T. Sandholm, "Superhuman ai for multiplayer poker," *Science*, 2019.

[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[26] P. S. Castro, S. Moitra, C. Gelada, S. Kumar, and M. G. Bellemare, "Dopamine: A research framework for deep reinforcement learning," 2018. arXiv: 1812.06110.

[27] E. Cetin, P. J. Ball, S. Roberts, and O. Celiktutan, "Stabilizing off-policy deep reinforcement learning from pixels," in *Proc. of the International Conf. on Machine learning (ICML)*, 2022.

[28] R. Chekroun, M. Toromanoff, S. Hornauer, and F. Moutarde, "GRI: general reinforced imitation and its application to vision-based autonomous driving," 2021. arXiv: 2111.08575.

[29] E. Chen, Z. Hong, J. Pajarinen, and P. Agrawal, "Redeeming intrinsic rewards via constrained optimization," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[30] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[31] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[32] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[33] J. Clark and D. Amodei, *Faulty reward functions in the wild*, URL: https://openai.com/research/faulty-reward-functions, 2016.

[34] K. Cobbe, J. Hilton, O. Klimov, and J. Schulman, "Phasic policy gradient," in *Proc. of the International Conf. on Machine learning (ICML)*, 2021.

[35] R. Coulom, "Efficient selectivity and backup operators in monte-carlo tree search," in *International Conference on Computers and Games (ICCG)*, 2006.

[36] W. Dabney, G. Ostrovski, D. Silver, and R. Munos, "Implicit quantile networks for distributional reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[37] W. Dabney, M. Rowland, M. G. Bellemare, and R. Munos, "Distributional reinforcement learning with quantile regression," in *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2018.

[38] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. D. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, C. Donner, L. Fritz, C. Galperti, A. Huber, J. Keeling, M. Tsimpoukelli, J. Kay, A. Merle, J. Moret, S. Noury, F. Pesamosca, D. Pfau, O. Sauter, C. Sommariva, S. Coda, B. Duval, A. Fasoli, P. Kohli, K. Kavukcuoglu, D. Hassabis, and M. A. Riedmiller, "Magnetic control of tokamak plasmas through deep reinforcement learning," *Nature*, 2022.

[39] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," 2012. arXiv: 1205.4839.

[40] J. Deng, W. Dong, R. Socher, L.-j. Li, K. Li, and L. Fei-fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[41] M. Dennis, N. Jaques, E. Vinitsky, A. M. Bayen, S. Russell, A. Critch, and S. Levine, "Emergent complexity and zero-shot transfer via unsupervised environment design," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[42] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, Y. Wu, and P. Zhokhov, *Openai baselines*, 2017. URL: https://github.com/openai/baselines.

[43] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester, "An empirical investigation of the challenges of real-world reinforcement learning," 2020. arXiv: 2003.11881.

[44] O. Eberhard, J. Hollenstein, C. Pinneri, and G. Martius, "Pink noise is all you need: Colored noise exploration in deep reinforcement learning," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.

[45] A. Ecoffet, J. Huizinga, J. Lehman, K. O. Stanley, and J. Clune, "First return, then explore," *Nature*, 2021.

[46] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, "Implementation matters in deep policy gradients: A case study on PPO and TRPO," *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

[47] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures," in *Proc. of the International Conf. on Machine learning (ICML)*, J. G. Dy and A. Krause, Eds., 2018.

[48] J. Fan, Z. Wang, Y. Xie, and Z. Yang, "A theoretical analysis of deep q-learning," in *Proceedings of the Conference on Learning for Dynamics and Control, (L4DC)*, 2020.

[49] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee, "DPOK: reinforcement learning for fine-tuning text-to-image diffusion models," 2023. arXiv: 2305.16381.

[50] A. Fawzi, M. Balog, A. Huang, T. Hubert, B. Romera-Paredes, M. Barekatain, A. Novikov, F. J. R. Ruiz, J. Schrittwieser, G. Swirszcz, D. Silver, D. Hassabis, and P. Kohli, "Discovering faster matrix multiplication algorithms with reinforcement learning," *Nature*, 2022.

[51] M. Fortunato, M. G. Azar, B. Piot, J. Menick, M. Hessel, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg, "Noisy networks for exploration," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018.

[52] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, "An introduction to deep reinforcement learning," *Foundations and Trends in Machine Learning*, 2018.

[53] P. I. Frazier, "A tutorial on bayesian optimization," 2018. arXiv: 1807.02811.

[54] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[55] R. Geirhos, J. Jacobsen, C. Michaelis, R. S. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, 2020.

[56] A. Gosavi, "Reinforcement learning: A tutorial survey and recent advances," *INFORMS J. Comput.*, 2009.

[57] J. Grabocka, R. Scholz, and L. Schmidt-Thieme, "Learning surrogate losses," 2019. arXiv: 1905.10108.

[58] E. Greensmith, P. L. Bartlett, and J. Baxter, "Variance reduction techniques for gradient estimates in reinforcement learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2001.

[59] E. Greensmith, P. L. Bartlett, and J. Baxter, "Variance reduction techniques for gradient estimates in reinforcement learning," *Journal of Machine Learning Research (JMLR)*, 2004.

[60] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. of the International Conf. on Machine learning (ICML)*, 2017.

[61] N. Gürtler, S. Blaes, P. Kolev, F. Widmaier, M. Wuthrich, S. Bauer, B. Schölkopf, and G. Martius, "Benchmarking offline reinforcement learning on real-robot hardware," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.

[62] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[63] 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. of the International Conf. on Machine learning (ICML)*, 2018.

[64] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018. arXiv: 1812.05905.

[65] D. Hafner, T. P. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

[66] D. Hafner, T. P. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

[67] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," 2023. arXiv: 2301.04104.

[68] N. Hansen, "The CMA evolution strategy: A tutorial," 2016. arXiv: 1604.00772.

[69] M. E. Harmon and S. S. Harmon, "Reinforcement learning: A tutorial," *WL/AAFC, WPAFB Ohio*, 1996.

[70] H. Hasselt, "Double q-learning," *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

[71] H. V. Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2016.

[72] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer, 2009.

[73] M. J. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2015.

[74] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[75] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep reinforcement learning that matters," in *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2018.

[76] M. Hessel, J. Modayil, H. V. Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Proc. of the Conf. on Artificial Intelligence (AAAI)*, 2018.

[77] D. Horgan, J. Quan, D. Budden, G. Barth-Maron, M. Hessel, H. V. Hasselt, and D. Silver, "Distributed prioritized experience replay," 2018. arXiv: 1803.00933.

[78] C. Huang, S. Zhai, P. Guo, and J. M. Susskind, "Metricopt: Learning to optimize black-box evaluation metrics," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[79] S. Huang, R. F. J. Dossa, A. Raffin, A. Kanervisto, and W. Wang, "The 37 implementation details of proximal policy optimization," in *ICLR Blog Track*, 2022. URL: https://iclr-blog-track.github.io/2022/03/25/ppo-implementation-details/.

[80] S. Huang, R. F. J. Dossa, C. Ye, J. Braga, D. Chakraborty, K. Mehta, and J. G. M. Araújo, "Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms," *Journal of Machine Learning Research (JMLR)*, 2022.

[81] M. Jiang, E. Grefenstette, and T. Rocktäschel, "Prioritized level replay," in *Proc. of the International Conf. on Machine learning (ICML)*, 2021.

[82] M. Jiang, T. Rocktäschel, and E. Grefenstette, "General intelligence requires rethinking exploration," *Royal Society Open Science*, 2023.

[83] D. J. Joshi, I. Kale, S. Gandewar, O. Korate, D. Patwari, and S. Patil, "Reinforcement learning: A survey," in *Machine Learning and Information Processing (ICMLIP)*, 2021.

[84] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research (JAIR)*, 1996.

[85] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski, "Model based reinforcement learning for atari," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

[86] S. Kapturowski, V. Campos, R. Jiang, N. Rakicevic, H. van Hasselt, C. Blundell, and A. P. Badia, "Human-level atari 200x faster," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023.

[87] S. Kapturowski, G. Ostrovski, J. Quan, R. Munos, and W. Dabney, "Recurrent experience replay in distributed reinforcement learning," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019.

[88] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, "Champion-level drone racing using deep reinforcement learning," *Nature*, 2023.

[89] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014.

[90] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1008–1014, 1999.

[91] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[92] K. A. Krueger and P. Dayan, "Flexible shaping: How learning in small steps helps," *Cognition*, 2009.

[93] A. Kumar, X. B. Peng, and S. Levine, "Reward-conditioned policies," 2019. arXiv: 1912.13465.

[94] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[95] A. Lazaridis, A. Fachantidis, and I. P. Vlahavas, "Deep reinforcement learning: A state-of-the-art walkthrough," *Journal of Artificial Intelligence Research (JAIR)*, 2020.

[96] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, 2020.

[97] J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel, "Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research," 2019. arXiv: 1903.00742.

[98] S. Levine, *Cs 285: Deep reinforcement learning*, URL: https://rail.eecs.berkeley.edu/deeprlcourse/, 2023.

[99] S. Levine, A. Kumar, G. Tucker, and J. Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," 2020. arXiv: 2005.01643.

[100] C. Li and L. D. Pyeatt, "A short tutorial on reinforcement learning," in *Intelligent Information Processing II*, 2004.

[101] M. Li, J. Zhang, and E. Bareinboim, "Causally aligned curriculum learning," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2024.

[102] Y. Li, "Deep reinforcement learning," 2018. arXiv: 1810.06339.

[103] X. Liang, T. Wang, L. Yang, and E. P. Xing, "CIRL: controllable imitative reinforcement learning for vision-based self-driving," in *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018.

[104] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2016.

[105] L. J. Lin, "Self-improving reactive agents based on reinforcement learning, planning and teaching," *Machine Learning*, 1992.

[106] X. Lu, B. V. Roy, V. Dwaracherla, M. Ibrahimi, I. Osband, and Z. Wen, "Reinforcement learning, bit by bit," *Foundations and Trends in Machine Learning*, 2023.

[107] D. J. Mankowitz, A. Michi, A. Zhernov, M. Gelmi, M. Selvi, C. Paduraru, E. Leurent, S. Iqbal, J.-B. Lespiau, A. Ahern, T. Köppe, K. Millikin, S. Gaffney, S. Elster, J. Broshear, C. Gamble, K. Milan, R. Tung, M. Hwang, T. Cemgil, M. Barekatain, Y. Li, A. Mandhane, T. Hubert, J. Schrittwieser, D. Hassabis, P. Kohli, M. Riedmiller, O. Vinyals, and D. Silver, "Faster sorting algorithms discovered using deep reinforcement learning," *Nature*, 2023.

[108] P. Marbach and J. N. Tsitsiklis, "Simulation-based optimization of markov reward processes," *IEEE Trans. on Automatic Control (TAC)*, 2001.

[109] N. Metropolis and S. Ulam, "The monte carlo method," *Journal of the American Statistical Association (JASA)*, 1949.

[110] N. Meuleau, L. Peshkin, and K.-E. Kim, "Exploration in gradient-based reinforcement learning," *Technical Report*, 2001.

[111] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science Robotics*, 2022.

[112] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2016.

[113] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, 2015.

[114] M. Moravčík, M. Schmid, N. Burch, V. Lisỳ, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling, "Deepstack: Expert-level artificial intelligence in heads-up no-limit poker," *Science*, 2017.

[115] S. S. Mousavi, M. Schukat, and E. Howley, "Deep reinforcement learning: An overview," in *Proceedings of Intelligent Systems Conference (IntelliSys)*, 2016.

[116] K. Narasimhan, T. D. Kulkarni, and R. Barzilay, "Language understanding for text-based games using deep reinforcement learning," in *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.

[117] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in *Proc. of the International Conf. on Machine learning (ICML)*, 1999.

[118] E. Nikishin, M. Schwarzer, P. D'Oro, P. Bacon, and A. C. Courville, "The primacy bias in deep reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2022.

[119] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[120] F. Pardo, A. Tavakoli, V. Levdik, and P. Kormushev, "Time limits in reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[121] J. Parker-Holder, M. Jiang, M. Dennis, M. Samvelyan, J. N. Foerster, E. Grefenstette, and T. Rocktäschel, "Evolving curricula with regret-based environment design," in *Proc. of the International Conf. on Machine learning (ICML)*, ser. Proceedings of Machine Learning Research, 2022.

[122] J. Peng and R. J. Williams, "Incremental multi-step q-learning," *Machine Learning*, 1996.

[123] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *Communications of the ACM*, 2018.

[124] J. Perolat, B. D. Vylder, D. Hennes, E. Tarassov, F. Strub, V. de Boer, P. Muller, J. T. Connor, N. Burch, T. Anthony, S. McAleer, R. Elie, S. H. Cen, Z. Wang, A. Gruslys, A. Malysheva, M. Khan, S. Ozair, F. Timbers, T. Pohlen, T. Eccles, M. Rowland, M. Lanctot, J.-B. Lespiau, B. Piot, S. Omidshafiei, E. Lockhart, L. Sifre, N. Beauguerlange, R. Munos, D. Silver, S. Singh, D. Hassabis, and K. Tuyls, "Mastering the game of stratego with model-free multiagent reinforcement learning," *Science*, 2022.

[125] A. Petrenko, Z. Huang, T. Kumar, G. S. Sukhatme, and V. Koltun, "Sample factory: Egocentric 3d control from pixels at 100000 FPS with asynchronous reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2020.

[126] A. S. Pinto, A. Kolesnikov, Y. Shi, L. Beyer, and X. Zhai, "Tuning computer vision models with task rewards," in *Proc. of the International Conf. on Machine learning (ICML)*, 2023.

[127] R. Portelas, C. Colas, L. Weng, K. Hofmann, and P. Oudeyer, "Automatic curriculum learning for deep RL: A short survey," in *Proc. of the International Joint Conf. on Artificial Intelligence (IJCAI)*, 2020.

[128] R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini, "A survey on offline reinforcement learning: Taxonomy, review, and open problems," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[129] 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," 2023. arXiv: 2305.18290.

[130] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research (JMLR)*, 2021.

[131] S. Ross, G. J. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[132] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, 1959.

[133] T. Schaul, A. Barreto, J. Quan, and G. Ostrovski, "The phenomenon of policy churn," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[134] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," 2015. arXiv: 1511.05952.

[135] J. Schmidhuber, "Reinforcement learning upside down: Don't predict rewards - just map them to actions," 2019. arXiv: 1912.02875.

[136] B. Schölkopf, "Causality for machine learning," in *Probabilistic and Causal Inference: The Works of Judea Pearl*, 2022.

[137] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. P. Lillicrap, and D. Silver, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, 2020.

[138] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. of the International Conf. on Machine learning (ICML)*, 2015.

[139] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2016.

[140] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. arXiv: 1707.06347.

[141] M. Schwarzer, J. S. Obando-Ceron, A. C. Courville, M. G. Bellemare, R. Agarwal, and P. S. Castro, "Bigger, better, faster: Human-level atari with human-level efficiency," in *Proc. of the International Conf. on Machine learning (ICML)*, ser. Proceedings of Machine Learning Research, 2023.

[142] O. G. Selfridge, R. S. Sutton, and A. G. Barto, "Training and tracking in robotics," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 1985.

[143] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy," *IEEE Trans. Inf. Theory*, 1980.

[144] D. Silver, *Lectures on reinforcement learning*, URL: https://www.davidsilver.uk/teaching/, 2015.

[145] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of go with deep neural networks and tree search," *Nature*, 2016.

[146] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, 2018.

[147] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. of the International Conf. on Machine learning (ICML)*, 2014.

[148] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. P. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, 2017.

[149] Y. Song, A. G. Schwing, R. S. Zemel, and R. Urtasun, "Training deep neural networks via direct loss minimization," in *Proc. of the International Conf. on Machine learning (ICML)*, 2016.

[150] R. K. Srivastava, P. Shyam, F. Mutz, W. Jaskowski, and J. Schmidhuber, "Training agents using upside-down reinforcement learning," 2019. arXiv: 1912.02877.

[151] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize with human feedback," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

[153] R. S. Sutton, D. A. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in Neural Information Processing Systems (NIPS)*, 1999.

[154] C. Szepesvári, *Algorithms for Reinforcement Learning*. Morgan & Claypool Publishers, 2010.

[155] G. Tesauro, "Temporal difference learning and td-gammon," *Communications of the ACM*, 1995.

[156] S. Thrun and A. Schwartz, "Issues in using function approximation for reinforcement learning," in *Proceedings of the 1993 connectionist models summer school*, 1993.

[157] M. Toromanoff, E. Wirbel, and F. Moutarde, "End-to-end model-free reinforcement learning for urban driving using implicit affordances," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[158] J. N. Tsitsiklis, "Asynchronous stochastic approximation and q-learning," *Machine Learning*, 1994.

[159] G. Tucker, S. Bhupatiraju, S. Gu, R. E. Turner, Z. Ghahramani, and S. Levine, "The mirage of action-dependent baselines in reinforcement learning," in *Proc. of the International Conf. on Machine learning (ICML)*, 2018.

[160] M. Vecerík, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. A. Riedmiller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," 2017. arXiv: 1707.08817.

[161] M. Vidyasagar, "A tutorial introduction to reinforcement learning," 2023. arXiv: 2304.00803.

[162] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, Ç. Gülçehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. P. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in starcraft II using multi-agent reinforcement learning," *Nature*, 2019.

[163] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik, "Diffusion model alignment using direct preference optimization," 2023. arXiv: 2311.12908.

[164] H. Wang, N. Liu, Y. Zhang, D. Feng, F. Huang, D. S. Li, and Y. Zhang, "Deep reinforcement learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, 2020.

[165] L. Wang, Q. Cai, Z. Yang, and Z. Wang, "Neural policy gradient methods: Global optimality and rates of convergence," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2020.

[166] X. Wang, S. Wang, X. Liang, D. Zhao, J. Huang, X. Xu, B. Dai, and Q. Miao, "Deep reinforcement learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[167] C. J. Watkins and P. Dayan, "Q-learning," *Machine Learning*, 1992.

[168] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.

[169] L. Weaver and N. Tao, "The optimal reward baseline for gradient-based reinforcement learning," in *UAI '01: Proc. of the Conference in Uncertainty in Artificial Intelligence*, 2001.

[170] T. V. de Wiele, D. Warde-Farley, A. Mnih, and V. Mnih, "Q-learning in enormous action spaces via amortized approximate maximization," 2020. arXiv: 2001.08116.

[171] R. Wightman, *Pytorch image models*, 2019. DOI: 10.5281/zenodo.4414861.

[172] R. J. Williams and J. Peng, "Function optimization using connectionist reinforcement learning algorithms," *Connection Science*, 1991.

[173] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, 1992.

[174] T. Wu, B. Zhu, R. Zhang, Z. Wen, K. Ramchandran, and J. Jiao, "Pairwise proximal policy optimization: Harnessing relative feedback for LLM alignment," 2023. arXiv: 2310.00212.

[175] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, P. Khandelwal, V. Kompella, H. Lin, P. MacAlpine, D. Oller, T. Seno, C. Sherstan, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Dürr, P. Stone, M. Spranger, and H. Kitano, "Outracing champion gran turismo drivers with deep reinforcement learning," *Nature*, 2022.

[176] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022.

[177] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in *Proc. of the International Conf. on Learning Representations (ICLR)*, 2021.

[178] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021.