A Review of Safe Reinforcement Learning: Methods, Theories and Applications

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, Alois Knoll, *Fellow, IEEE*

***Abstract*—Reinforcement Learning (RL) has achieved tremendous success in many complex decision-making tasks. However, safety concerns are raised during deploying RL in real-world applications, leading to a growing demand for safe RL algorithms, such as in autonomous driving and robotics scenarios. While safe control has a long history, the study of safe RL algorithms is still in the early stages. To establish a good foundation for future safe RL research, in this paper, we provide a review of safe RL from the perspectives of methods, theories, and applications. Firstly, we review the progress of safe RL from five dimensions and come up with five crucial problems for safe RL being deployed in real-world applications, coined as "2H3W". Secondly, we analyze the algorithm and theory progress from the perspectives of answering the "2H3W" problems. Particularly, the sample complexity of safe RL algorithms is reviewed and discussed, followed by an introduction to the applications and benchmarks of safe RL algorithms. Finally, we open the discussion of the challenging problems in safe RL, hoping to inspire future research on this thread. To advance the study of safe RL algorithms, we release an open-sourced repository containing the implementations of major safe RL algorithms at the link[1].**

***Index Terms*—Safe reinforcement learning, safety optimisation, constrained Markov decision processes, safety problems.**

I. Introduction

OVER the past decades, Reinforcement Learning (RL) has been widely adopted in many fields, e.g. transportation schedule [1], recommender systems [2], and robotics [3], etc. However, a challenging problem in this domain is: **how do we guarantee safety when we deploy RL in real-world applications?** After all, unacceptable catastrophes may arise if we fail to take safety into account during RL applications in real-world scenarios. For example, it must not hurt humans when robots interact with humans in human-machine interaction environments; false or racially discriminating information should not be recommended for people in recommender systems; safety has to be ensured when self-driving cars are carrying out tasks in real-world environments. More specifically, we introduce several safety definitions from different perspectives, which might be helpful for safe RL research.

Shangding Gu, Florian Walter and Alois Knoll are with the Department of Informatics, Technical University of Munich, Munich 85748, Germany (e-mail: shangding.gu@tum.de; florian.walter@tum.de; knoll@mytum.de).

Long Yang is with the Institute for AI, Peking University, Beijing 100871, China (e-mail: yanglong001@pku.edu.cn).

Yali Du is with the Department of Informatics, King's College London, London, WC1E 6EB, UK (e-mail: yali.du@kcl.ac.uk).

Guang Chen is with the Department of Automotive Engineering, Tongji University, Shanghai 201804, China (e-mail: guangchen@tongji.edu.cn).

Jun Wang is with the Department of Computer Science, University College London, London WC1E 6BT, UK (e-mail: jun.wang@cs.ucl.ac.uk).

[1]https://github.com/chauncygu/Safe-Reinforcement-Learning-Baselines.git

Safety definition. The first type of safety definition: according to the definition of Oxford dictionary [4], the phrase "safety" is commonly interpreted to mean "the condition of being protected from or unlikely to cause danger, risk, or injury." The second type of safety definition: the definition of general "safety" according to wiki [2], the state of being "safe" is defined as "being protected from harm or other dangers"; "controlling recognized dangers to attain an acceptable level of risk" is also referred to as "safety". The third type of safety definition: according to Hans *et al.* [5], humans need to label environmental states as "safe" or "unsafe," and agents are considered "safe" if "they never reach unsafe states". The fourth type of safety definition: agents are considered to be "safe" by some research [6]–[8] if "they act, reason, and generalize obeying human desires". The fifth type of safety definitions: Moldovan and Abbeel [9] consider an agent "safe" if "it meets an ergodicity requirement: it can reach each state it visits from any other state it visits, allowing for reversible errors". In this review, based on the above various definitions, we investigate safe RL methods, which are about optimizing cost objectives, avoiding adversary attacks, improving undesirable situations, reducing risk, and controlling agents to be safe, etc.

RL safety is a significant practical problem that confronts us in RL applications, and is one of the critical problems in AI safety [10] that remains unsolved, though it has attracted increasing attention in the field of RL. Moreover, it is deduced that the mean minus variance [11] and percentile optimisation [12] of safe RL are, in general, NP-hard problems [13]. In some applications, the agent's safety is much more important than the agent's reward [14]. An attempt to answer the question above raises some fundamental problems that we call **"2H3W"** problems:

(1) **Safety Policy. H**ow can we perform policy optimisation to search for a safe policy?
(2) **Safety Complexity. H**ow much training data is required to find a safe policy?
(3) **Safety Applications. W**hat is the up to date progress of safe RL applications?
(4) **Safety Benchmarks. W**hat benchmarks can we use to fairly and holistically examine safe RL performance?
(5) **Safety Challenges. W**hat are the challenges faced in future safe RL research?

Most of the research in this field is aimed at solving the above **"2H3W"** problems, and the framework of safe RL about **"2H3W"** problems is shown in Figure 1.

[2]https://en.wikipedia.org/wiki/Safety

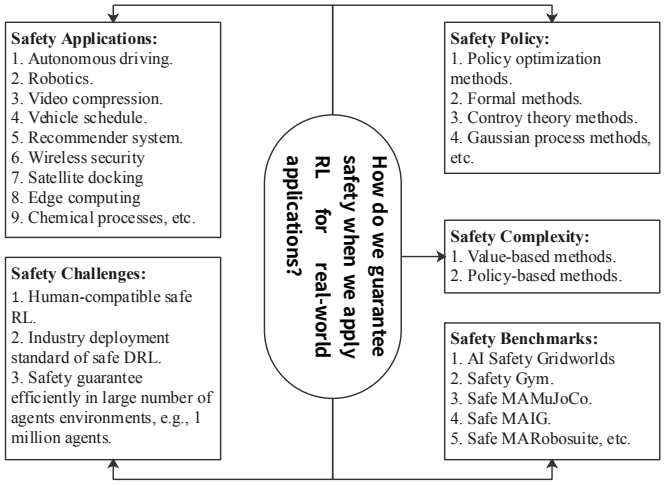


Fig. 1: The framework of safe RL about **"2H3W"** problems.

As for the problem (1) (**safety policy**), in many practical applications, a robot must not visit some states, and must not take some actions, which can be thought of as "unsafe" either for itself or for elements of its environment. It is essential that a safe policy function or value function is provided so that agents can reach safe states or perform safe actions. To achieve safety policy, a growing number of approaches have been developed over the last few decades, such as policy optimization methods [3], [15]–[27], formal methods [28]–[33], controy theory methods [34]–[38], Gaussian processes methods [39]–[47].

As for problem (2) (**safety complexity**), agents need to interact with environments and sample trajectories, such that the safe RL algorithms converge to below the constraint bound, and guarantee application safety. Since one of the natures of RL is exploration learning [48], it is usually hard to control the balance between exploitation and exploration, especially when we need to improve reward performance while satisfying cost constraints (cost is one way of encoding safety). Thus, we need to determine how many sample trajectories can make safe RL algorithms converge and satisfy safety bounds using the analysis of safety complexity during safe RL applications. Furthermore, the sample complexity of each safe RL algorithm is investigated from the viewpoints of value-based [49]–[52], and policy-based [53]–[55] methods, etc.

As for problem (3) (**safety applications**), although there are many RL applications to date, most of the applications are merely simulations that do not take safety into account; some real-world experiments have been carried out, but there is still a long way to go before RL can be used in real-world applications. Generally, when we use safe RL in real-world applications, we need to consider the ego agent safety, environmental safety, and human safety. Furthermore, we need to consider the control safety, which prevents adversary attacks from destroying or controlling the agent [56]. Therefore, for the safe RL application research, we introduce some safe RL applications in this review, e.g., autonomous driving [57]–[64], robotics [65]–[70], video compression [71], vehicle schedule [1], [72], etc.

As for problem (4) (**safety benchmarks**), we need to determine how to design cost and reward functions considering the balance between RL reward performance and safety in each benchmark, since cost functions will typically disturb reward performance. If we have a loose cost function, we may not be able to guarantee agent safety during the learning process; if we take too conservative cost functions, for example, in a constrained policy optimization process, when we set the cost constraint bound as zero or a negative value, which may result in lousy reward performance. Thus, we should pay more attention to designing the cost and reward function in the benchmarks. In some safe RL benchmarks, such as AI Safety Gridworlds [73], Safety Gym [74], Safe MAMuJoCo [75], Safe MAIG [3], Safe MARobosuite [75], the cost and reward functions are tailored to specific tasks well in examining experiments.

As for problem (5) (**safety challenges**), firstly, when we consider RL safety, it is a significant challenge about how to consider human safety factors or environmental safety factors during deploying RL in real-world applications. Secondly, a further important aspect when considering RL safety is how to take robot safety factors into account during RL applications [10], [56]. Another critical challenge is the social dilemma problem with safety balance. For example, the game of the trolley problem [76]. In the game, an out-of-control trolley will eventually kill five people if no action is taken, but you can redirect the trolley to another track, where only one person will be killed. The open question is how to balance safety weights when using RL. Moreover, the application standard and safe Multi-Agent RL (MARL) should be considered for future research.

The problem (5) (**safety challenges**) appear to be dilemma problems. They are not more straightforward problems compared to problem (1) (**safety policy**), problem (2) (**safety complexity**), problem (3) (**safety applications**), and the problem (4)(**safety benchmarks**). We must guarantee agent, human, and environmental safety when we apply RL for practical applications by providing sophisticated algorithms. In addition, few studies have focused on problem (2) (**safety complexity**) and problem (4) (**safety benchmarks**), especially in industrial use. Answering problem (3) (**safety applications**) and problem (5) (**safety challenges**) may reveal RL application situations and provide a clue for future RL research. In this review, we will summarise the progress of safe RL, answer the five problems, and analyze safe RL algorithms, theory, and applications. In general, we mainly sort out the safe RL research of the past two decades (Though we are unfortunately unable to include some impressive safe RL literature in this review for space reasons).

The main contributions of this paper: first, we investigate safe RL research and give an indication of the research progress. Second, the main practical question of RL applications is discussed, and five fundamental problems are analyzed in detail. Third, algorithms, theory, and applications of safe RL are reviewed in detail, e.g., safe model-based learning and safe model-free learning, in which we present a bird's eye view to summarising the progress of safe RL. Finally, the challenges

[3]https://github.com/chauncygu/Safe-Multi-Agent-Isaac-Gym.git

we face when using RL for applications are explained. Due to page limit, this paper is a distilled essence of the preprint version [77], the more detail of the paper can be found in the preprint version [77].

II. Background

Safe RL is often modeled as a Constrained Markov Decision Process (CMDP) [78], in which we need to maximize the agent reward while making agents satisfy safety constraints. A substantial body of literature has studied CMDP problems for both tabular and linear cases [78]–[82]. However, deep safe RL for high dimensional and continuous CMDP optimization problems is a relatively new area that has emerged in recent years, and proximal optimal values generally represent safe states or actions using neural networks. In this section, we illustrate the generally deep safe RL problem formulation concerning the objective functions of safe RL and offer an introduction to safe RL surveys.

A. Problem Formulation of Safe Reinforcement Learning

A CMDP problem [78] is an extension of a standard Markov decision process (MDP) $\mathcal{M}$ with a constraint set $\mathcal{C}$. A tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \rho_0, \gamma)$ is given to present a MDP [83]. A state set is denoted as $\mathcal{S}$, an action set is denoted as $\mathcal{A}$, $\mathbb{P}(s^{'}|s, a)$ denotes the probability of state transition from s to s' after playing a. A reward function is denoted as $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. $\rho_0(\cdot) : \mathcal{S} \to [0, 1]$ is the starting state distribution, γ denotes the discount factor.

In safe RL, the general goal of an optimal policy π is to maximize the reward objective $J(\pi) = \mathbb{E}_{\pi, s_0 \sim \rho_0(\cdot)}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0 = s]$ and minimize the cost i's objective $C_i(\pi) = \mathbb{E}_{\pi, s_0 \sim \rho_0(\cdot)}[\sum_{t=0}^{\infty} \gamma^t c_{t+1} | s_0 = s]$ to below the safety constraint bound b_i by selecting an action a at time step t, and c denotes the cost value of each step, $\Pi_{\mathcal{C}}$ is the policy set, τ is a trajectory, $\tau = (s_0, a_0, s_1, \ldots)$, in which action depends on π, $s_0 \sim \rho_0(\cdot)$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim \mathbb{P}(\cdot|s_t, a_t)$. $d_\pi^{s_0}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}_\pi(s_t = s | s_0)$ denotes the state distribution (starting at s_0), the discounted state distribution based on the initial distribution $\rho_0(\cdot)$ is present as $d_\pi^{\rho_0}(s) = \mathbb{E}_{s_0 \sim \rho_0(\cdot)}[d_\pi^{s_0}(s)]$. Due to the page limit, the detail of problem formulation for safe RL is provided in the preprint version of our investigations [77].

B. A Survey of Safe Reinforcement Learning related surveys

Several surveys have already investigated safe RL problems and methods, e.g., [84]–[87]. However, the surveys do not provide comprehensive theoretical analysis for safe RL, such as sample complexity, nor do they focus on the critical problems of safe RL that we point out, **"2H3W"** problems. For example, [84] investigates safe RL from the perspective of control theory and robotics; the soft constraints, probabilistic constraints, and hard constraints are defined in the survey. Furthermore, they reviewed a large number of papers that are related to control theory and analyzed how to use control theory to guarantee RL safety and stability, such as using model predictive control [88], adaptive control [89], robust control [90], Lyapunov functions [36] for RL stability and safety. In the survey [85], they focus more on reviewing safe RL methods up to 2015. They categorize safe RL methods into two types: one is based on the safety of optimization criterion, where the worst case, risk-sensitive criterion, constrained criterion, etc., are taken into account to ensure safety. Another one is based on external knowledge or risk metric. In general, the external knowledge or risk metric is leveraged to guide the optimisation of RL safety. In contrast to the survey [85], the survey [86] focuses more on the techniques of safe learning, including MDP and non-MDP methods, such as RL, active learning, evolutionary learning. The survey [87] summarises safe model-free RL methods based on two kinds of constraints, namely cumulative constraints and instantaneous constraints.

As for safe RL methods of cumulative constraints, three types of cumulative constraints are introduced, which are a discounted cumulative constraint (Equation (1)), a mean valued constraint (Equation (2)), and a probabilistic constraint (Equation (3)), respectively.

$$C_i(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t C_i\left(s_t, a_t, s_{t+1}\right)\right] \leq b_i, \tag{1}$$

$$C_i(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\frac{1}{T}\sum_{t=0}^{T-1} C_i\left(s_t, a_t, s_{t+1}\right)\right] \leq b_i, \tag{2}$$

$$C_i(\pi) = P\left(\sum_{t} C_i\left(s_t, a_t, s_{t+1}\right) \leq b_i\right) \geq \eta, \tag{3}$$

where η denotes the safety probability.

When it comes to instantaneous constraints (Instantaneous constraints are those that apply immediately, with constraint costs occurring without delay. It is essential to consider these constraints at each step, rather than focusing solely on the cumulative constraints over a trajectory with T steps), instantaneous explicit and implicit constraints are given. The explicit ones have an accurate, closed-form expression that can be numerically checked, e.g., the cost that an agent generates during each step; the implicit does not have an accurate closed-form expression, e.g., the probability that an agent will crash into unsafe areas during each step. Although based on our investigation, most CMDP methods are based on cumulative cost optimization, a few CMDP methods focus on the immediate costs to optimize performance [91], and it is natural to take the cost of a whole trajectory rather than a state or action in some real-world applications, such as robot-motion planning and resource allocation [35]. Compared to the related surveys [84]–[87], our survey pays more attention to answering the **"2H3W"** problems, and provides safe RL algorithm analysis, sample complexity analysis and convergence investigation from the perspectives of model-based and model-free RL.

III. Methods of Safe Reinforcement Learning

There are several types of safe RL methods based on different criteria. For example, from the perspective of objective optimization, several methods consider cost as one of the optimization objectives to achieve safety, e.g., [92]–[100]. From the perspective of knowledge utilization, some methods

consider safety in the RL exploration process by leveraging external knowledge, e.g., [9], [35], [101]–[105]. From the perspectives of policy and value-based methods for safe RL, summarise as follows, policy-based safe RL: [3], [27], [53]–[55], [75], [106], value-based safe RL: [49]–[52], [107]–[110]. From the perspective of the agent number, we have safe RL methods in a safe single-agent RL setting and a safe multi-agent RL setting. More specifically, numerous safe RL methods are about the single-agent setting. An agent needs to explore the environments to improve its reward while keeping costs below the constraint bounds. In contrast to safe single-agent RL methods, safe multi-agent RL methods not only need to consider the ego agent's reward and other agent's reward, but also have to take into account the ego agent's safety and other agents' safety in an unstable multi-agent system.

In this section, we provide a concise but holistic overview of safe RL methods from a bird's eye view and attempt to answer the **"Safety Policy"** problem. Especially, we will introduce safe RL methods from the perspectives of model-based methods and model-free methods, in which safe single-agent RL and multi-agent RL will be analyzed in detail. In particular, we will introduce policy optimization-based approaches, formal methods-based approaches, control theory-based approaches, and Gaussian processes-based approaches in model-based and model-free settings, respectively. In addition, the model-based and model-free safe RL analysis are summarised in the preprint version of the paper [77].

A. Methods of Model-Based Safe Reinforcement Learning

Although accurate models are challenging to build and many applications lack models, model-based Deep RL (DRL) methods usually have a better learning efficiency than model-free DRL methods. There are still many scenarios for which we can apply model-based DRL methods, such as robotics, transportation planning, logistics, etc. Several works have shown that safe problems, such as a safe robot control problem [111], can be overcome by using model-based safe RL methods. In this section, we will investigate model-based safe RL methods from different perspectives: policy optimization-based approaches, control theory-based approaches, formal methods-based approaches, and Gaussian process-based approaches.

1) Policy optimization-based approaches: Policy optimization-based approaches usually search for a safe policy using cumulative cost values on trajectories. For example, Moldovan and Abbeel use the Chernoff function (4) in [13] to achieve near-optimal bounds with desirable theoretical properties. The cumulative expectation cost is represented in the function as $E_{s,\pi}\left[e^{J/\theta}\right]$, especially for δ, which can be used to adjust the balance of reward performance and safety. For instance, if δ is set as 1, this method will ignore the safety, and if δ is set as 0, this method will fully consider the risk and optimize the cost. Moreover, they examine the method in grid world environments and air travel planning applications. However, the method needs a significant amount of time to recover policy from risk areas, which is ten times the value iteration.

$$C^{\delta}_{s,\pi}[J] = \inf_{\theta>0}\left(\theta \log E_{s,\pi}\left[e^{J/\theta}\right] - \theta\log(\delta)\right) \tag{4}$$

Different from the Chernoff function based methods [13], Borkar [112] proposes an actor-critic RL method to handle a CMDP problem based on the envelope theorem in mathematical economics [113], [114], in which the primal-dual optimization is analyzed in detail using a three-time scale process. The critic scheme, actor scheme, and dual ascent are on the fast, middle, and slow timescale. Bharadhwaj *et al.* [23] also present an actor-critic method to address a safe RL problem, where they first develop a conservative safety critic to estimate the safety. The primal-dual gradient descent is leveraged to optimize the reward and cost value by constraining the failure probability. Although this method can bound the probability of failures during policy evaluation and improvement, this method still cannot guarantee total safety; a few unsafe corner actions may dramatically damage critical robot applications.

Akin to Borkar's method [112], Tessler *et al.* [21] also utilize a multi-timescale approach with regards to cost as a part of the reward in primal-dual methods. However, the method's learning rate is hard to tune for real-world applications because of imposing stringent requirements, and the method may not guarantee safety when agents are training.

Similar to the above methods, in actor-critic settings, with primal-dual optimization techniques, Yu *et al.* [115] convert a non-convex constrained problem into a locally convex problem and guarantee the stationary point to the optimal point of non-convex optimization problem; they consider the state-action safety optimization, and the optimization process is motivated by [116], in which the Lipschitz condition is necessary to satisfy the results. Also, they need to estimate the policy gradient for optimization.

Analogously, using optimization theory, the policy gradient and actor-critic methods are proposed by Chow *et al.* [117] to optimize risk RL performance, in which CVaR[4]-constrained and chance-constrained optimization are used to guarantee safety. Specifically, the importance sampling [118], [119] is used to improve policy estimation and provide convergence proof for the proposed algorithms. Nevertheless, this method may not guarantee safety during training [120]. Paternain *et al.* [18] provide a duality theory for CMDP optimization, and they prove the zero duality gap in primal-dual optimization even for non-convexity problems. Furthermore, they point out that the primal problem can be exactly solved by dual optimization. In their study, the suboptimal bound using neural network parametrization policy is also present [121].

2) Control theory-based approaches: Control theory-based approaches mostly ensure more rigorous safety than policy optimization approaches. This is because control theory-based methods typically utilize a physical model to regulate actions or states, whereas policy optimization methods often rely on data-driven approaches to learn safe actions or states. Consequently, model-driven methods tend to provide stronger safety assurances compared to their data-driven counterparts. This point is also supported from the work [84]. Nevertheless, control theory-based approaches may not be easy to transfer to other domains due to the strict requirements of dynamics models. For instance, Berkenkamp *et al.* [122] develop a safe

[4]CVaR denotes the Conditional Value at Risk.

model-based RL algorithm by leveraging Lyapunov functions to guarantee stability with the assumptions of Gaussian process prior; their method can ensure a high-probability safe policy for an agent in a continuous process. However, Lyapunov functions are usually hand-crafted, and it is not easy to find a principle to construct Lyapunov functions for an agent's safety and performance [35]. In addition, some safe RL methods are proposed from the perspective of Model Predictive Control (MPC) [123], e.g., MPC is used to make robust decisions in CMDPs [124] by leveraging a constrained MPC method [125], which also introduces a general safety framework to make decisions [126].

Apart from Lyapunov functions and MPC-safe RL approaches, control barrier functions (CBFs) based approaches are also proposed to ensure learning safety [34], [127]–[129]. Similar to the above control methods, CBFs based safe RL approaches also require a dynamics model for the control system. Thus, it is not straightforward to deploy it in RL [130]. For example, [127] propose a model-based safe RL by leveraging lagrangian optimization and CBFs; they further deploy their method in simulation and real-world autonomous driving experiments, the experiment results indicate that their method can improve sample efficiency and ensure safety. In contrast to [127], [128] not just use CBFs, but also robust CBFs [131], they develop a safe RL method based on SAC [132] and robust CBFs. Specifically, they consider safety in robust settings by estimating the disturbed dynamics systems with GP models. [129] propose a safe exploration framework based on model-based RL and CBFs, where Lyapunov-like CBFs [133] are used to construct the safety CBFs, and they can ensure the learning safety and stability. The above methods show impressive performance in ensuring safety. However, designing CBFs necessitates knowledge of the model or safety certificates, which may be challenging in complex environments, even when the model is derived using alternative methods.

3) Formal methods-based approaches: Different from policy optimization-based approaches, control theory-based approaches, and Gaussian Process-based methods, formal methods [28] usually try to ensure safety without unsafe probabilities. However, most formal methods rely heavily on the model knowledge and may not show better reward performance than other methods. The verification computation might be expensive for each neural network [28]. More generally, the curse of dimensionality problem is challenging to be solved, which appears when formal methods are deployed for RL safety [29], since formal methods may be intractable to verify RL safety in continuous and high-dimension space settings [29].

For instance, in [28], Anderson *et al.* provide a neurosymbolic RL method by leveraging formal verification to guarantee RL safety in a mirror descent framework. In their method, the neurosymbolic and constrained classes using symbolic policies are leveraged to approximate gradient and conduct verification in a loop-iteration setting. The experiment results show promising performance in RL safety, though the fixed worst-case model knowledge is used in these environments, which may not be suitable for practical applications. Similarly, in [29], Beard and Baheri utilize formal methods to improve agent safety by incorporating external knowledge and penalizing behaviors for RL exploration. Nonetheless, the methods may need to be developed for scalability and continuous systems. Finally, in [30], Fulton and Platzer also use external knowledge to ensure agent safety by leveraging the justified speculative control sandbox in offline formal verification settings.

4) Gaussian processes-based approaches: In formal methods, determining how to measure unsafe areas is a challenging problem. Recently, many approaches have been proposed by leveraging a Gaussian Process (GP) [134] to estimate the uncertainty and unsafe areas. Further, the information from GP methods is incorporated into the learning process for agent safety. For instance, Akametalu *et al.* [39] develop a safe RL method based on reachability analysis, in which they use GP methods to measure the disturbances which may lead to unsafe states for agents, the maximal safe areas are computed iteratively in an unknown dynamics system. Like Akametalu *et al.* [39] method, Berkenkamp and Schoellig [135] utilize GP methods to measure the system uncertainty and further guarantee system stability. Polymenakos *et al.* [43]develop a safe policy search approach based on PILCO (Probabilistic Inference for Learning Control) method [136] which is a policy gradient method derived from a GP, in which they improve agent safety using probability trajectory predictions by incorporating cost into reward functions. Similar to Polymenakos *et al.* [43], Cowen *et al.* [41] also use PILCO to actively explore environments while considering risk, a GP is used to quantify the uncertainty during exploration, and agent safety probability is improved by leveraging a policy multi-gradient solver.

The above safe RL methods present excellent performance in terms of the balance between reward and safety performance in most challenging tasks. Nonetheless, training safety or stability may need to be further investigated rigorously, and a unified framework may need to be proposed to better examine safe RL performance.

B. Methods of Model-Free Safe Reinforcement Learning

Most studies pay attention to model-free safe RL since it can be deployed in many domains without requiring model dynamics. In this section, we also investigate safe RL methods from the perspectives of policy optimization, control theory, Gaussian processes, and formal methods.

1) Policy optimization-based approaches: Constrained Policy Optimisation (CPO) [3] is the first policy gradient method to solve the CMDP problem based on model-free deep RL. In particular, a policy has to be optimized to guarantee the reward of a monotonic improvement while satisfying safety constraints. As a result, their methods can almost converge to safety bound and produce more comparable performance than the primal-dual method [117] on some tasks. However, CPO's computation is more expensive than PPO[5]-Lagrangian, since it needs to compute the Fisher information matrix and uses the second Taylor expansion to optimize objectives. Moreover, the approximation and sampling errors may have detrimental effects on the overall performance, and the convergence analysis is challenging. Furthermore, the additional recovery policy

[5]PPO denotes the Policy Proximal Optimization.

may require more samples, which could result in wasted samples [120].

Derived from CPO [3], Projection-based Constrained Policy Optimisation (PCPO) [27] based on two-step methods constructs a cost projection to optimize cost and guarantee safety, which displays better performance than CPO on some tasks. PCPO leverages policy to maximize the reward via Trust Region Policy Optimization (TRPO) method [137], and then projects the policy to a feasible region to satisfy safety constraints. However, the second-order proximal optimization is used in both steps, which may result in a more expensive computation than the First Order Constrained Optimization in Policy Space (FOCOPS) method [120], which only uses the first-order optimization. [120] is motivated by the optimization-based idea [138], where they use the primal-dual method, address policy search in the nonparametric space and project the policy into the parameter space, to carry out proximal maximization optimization in CMDPs. Although this method is easy to implement and shows better sample efficiency, it still needs to solve the problems of unstable saddle points and unsafe actions during training.

Similar to CPO [3], in Safe Advantage-based Intervention for Learning policies with Reinforcement (SAILR) [139], Wagener *et al.* leverage the advantage function as a surrogate to minimize cost, and further achieve safe policy both during training and deployment. Furthermore, In [140], a Shortest-Path Reinforcement Learning (SPRL) method is proposed using off-policy strategies to construct safe policy and reward policy, and its applications are used for the shortest path problems in Travel Sale Path (TSP). Besides, based on a Gaussian process of a safe RL method [47], the SNO-MDP[6] [141] is developed to optimize cost in the safe region and optimize the reward in the unknown safety-constrained region [45], [46]. It is suggested that the maximization reward is more important than safety. The policy is often substantial in some cases. For example, staying in the current position for safety is extremely conservative. This method [141] shows the near-optimal cumulative reward under some assumptions, whereas it cannot achieve the near-optimal values while guaranteeing safety constraints.

Different from CPO, a first-order policy optimization method [142] is provided based on interior point optimization (IPO) [143], in which the logarithmic barrier functions are leveraged to satisfy safety constraints. The method is easy to implement by tuning the penalty function. Although the empirical results on MuJoCo [3] and grid-world environments [144] have demonstrated their method's effectiveness, the theoretical analysis to guarantee the performance is still needed to be provided.

Unlike the above CMDP optimization, a meta algorithm [52] is proposed to solve safe RL problems with general convex constraints [145]. They can, in particular, solve CMDP problems with a small number of samples in reward-free settings. The algorithms can also be used to solve approachability problems [146], [147]. Although the theoretical results have shown their method's effectiveness, practical algorithms and experiments ought to be proposed and carried out to further evaluate their method. In [148], an attempt is made to solve safe RL using the trajectory probability in a safe set. The probability invariance is positive, and constraint gradients are obtained using the policy parameters. More importantly, the related problem can have an arbitrarily small duality gap. However, this method may also encounter the stability problems of primal-dual methods, and the saddle points using Lagrangian methods may be unstable [149].

Comparable to the above primal-dual settings, Ghosh *et al.* [150] have developed a model of safe reinforcement learning based on least-squares value iteration-upper confidence bound [151] in primal-dual settings, which is suitable for large-scale optimization, and the regret bound and safety violation are analyzed. While similar methodologies have been proposed, they are predominantly utilized within tabular settings, as exemplified by the work of Wei et al. [110]. Furthermore, Xiong *et al.* [152] introduce a step-wise safe RL method based on upper confidence bound value iteration [153]. This method employs optimistic estimations of transition kernels and cost functions, thereby enhancing its capacity to ensure safety in reward-free scenarios. Notably, Xiong *et al.* also provide analyses of safety violations and regret bounds, facilitating the extension of this method to additional contexts such as multi-robot systems. Although these methods demonstrate improved performance over related baselines, further exploration into their applications across diverse domains, such as safe multi-robot control and planning, could yield even more beneficial insights.

2) Control theory-based approaches: In contrast to safe policy optimization such as meta algorithms, CPO and IPO based on model-free RL, control theory is are also leveraged to ensure RL safety. For example, the first Lyapunov functions used for a safe RL may be [67], in which the agent's actions are constrained by applying the control law of Lyapunov functions to learning systems and removing the unsafe actions in the action set. The experiments have demonstrated that their method can achieve safe actions for the control problems in their study. However, the method requires the knowledge of a Lyapunov function in advance. If the environment dynamics model is unknown, it may be difficult to address safe RL problems with this method. Unlike [67], in [35] and [36], Chow *et al.* propose several safe RL methods based on Lyapunov functions in discrete and continuous CMDP settings, respectively, where Lyapunov functions are used for safe RL to guarantee safe policy and learning stability. The methods can guarantee safety during training, and the Lyapunov functions can be designed by a proposed Linear Programming algorithm. However, the training stability and safety using Lyapunov functions still need to be improved, and more efficient algorithms in the setting may need to be proposed.

Apart from Lyapunov functions based safe RL, CBFs are also developed in model-free settings in recent years [154], [155], [154] develop a model-free safe RL method based CBFs, in which they learn the policy and CBFs with data-driven methods, which can help to reduce reliance on the environment models. Similarly, [155] proposes a safe RL method for power system control with CBFs, in which they integrated CBFs into reward functions to search for safe policy. model free control

[6]The SNO-MDP represents the Safe Near-Optimal MDP.

barrier function, [156], [157]. Moreover, [158] leverages a model-free RL and CBFs to guarantee learning safety while improving learning efficiency, in which CBFs are used to constrain the search space to enhance learning safety and efficiency based on a GP learning model. Moreover, [130] also leverages CBFs to ensure learning safety with learned dynamics knowledge, where they provide convergence and stability analysis, and their method outperforms the baselines in their experiments. A promising approach to enhancing CBFs for real-world safe RL applications involves learning CBFs using neural networks, as demonstrated in safe control methods [154], [159], [160]. This technique leverages neural networks to model CBFs, offering significant advantages in complex scenarios. By employing neural networks, we can use the neural CBFs for safety guarantees more effectively, potentially leading to improved performance and broader applicability of safe RL in challenging real-world applications.

Safety layer methods have been proposed in recent years, which are more directly to ensure RL safety than safe policy optimization methods that based on cost value optimization such as CPO and IPO. We categorize safety-layer-based approaches into control theory methods since they can be considered as a control filter in control theory.

For example, Pham *et al.* propose an OptLayer [68] method, in which they leverage stochastic control policies to attain the reward performance, and a layer of the neural network is integrated to pursue safety during applications. The real-world applications also demonstrate the effectiveness of the safety layer. Qin *et al.* [161] propose the DCRL (Density Constrained Reinforcement Learning) method that is to optimize the reward and cost from the perspective of an optimization criterion, in which they consider the safety constraints via the duality property [162]–[165] with regard to the state density functions, rather than the cost functions, reward functions and value functions [3], [53], [166]. This method lies in model-free settings whereas Chen *et al.* [167] provide similar methods in model-based settings. A-CRL (state Augmented Constrained Reinforcement Learning) method [168] is proposed to address a CMDP problem whereby the optimal policy may not be achieved via regular rewards. Their method focuses on solving the monitor problem in CMDP while the dual gradient descent is used to find the feasible trajectories and guarantee safety. Nonetheless, OptLayer [68], A-CRL [168] and DCRL [161] all lack convergence rate analysis.

3) Gaussian processes-based approaches: Relative to safe policy optimization methods that attend more to cost values, Lyapunov function-based techniques, which emphasize safe actions, and Safety layer approaches, GP methodologies primarily concentrate on the safe exploration via modeling the safe states. In a GP of model-free settings, Sui *et al.* [45] use a GP method to present the unknown reward function from noise samples, the exploration by leveraging a GP method is improved to reduce uncertainty and ensure agent safety. More particularly, the GP method is used to predict unknown function, and guide the exploration in bandit settings which do not need state transitions. Their real-world experiments in movie recommendation systems and medical areas indicate that their method can achieve near-optimal values safely. Like Sui *et al.* [45], Turchetta *et al.* [46] also leverage a GP method to approximate unknown functions prior for safe exploration. Nevertheless, they focus more on finite MDP settings considering explicitly reachability. Nonetheless, the method may not optimize reward objectives while considering safety. Wachi *et al.* [47] represent unknown reward and cost functions with GP methods to ensure safety with probability and optimize reward. Furthermore, the safe, unsafe, and uncertain states are denoted for agent optimistic and pessimistic exploration, and their method can adapt the trade-off between exploration and exploitation. Nevertheless, the convergence guarantee with finite-time rates, optimization for multiple and heterogeneous objectives may need to be provided.

Although GP methods have shown impressive performance with regard to RL safety, most of them ensure safety with probability. How to rigorously guarantee RL safety during exploration still remains open.

4) Formal methods-based approaches: Distinct from employing GP methods to model the safe state, Hasanbeig et al. [169] propose a safety-enhanced RL approach that leverages LTL. In this novel methodology, logical formulas serve as constraints during the exploration phase of policy synthesis. Although this method demonstrates notable safety performance, it is imperative to carefully define the logical constraints to ensure safe exploration and effectively balance the trade-off between safety performance and reward acquisition. Murugesan *et al.* [170] employ a formal tool, satisfiability modulo theories [171], as a safety verification layer to enhance learning safety. Furthermore, Hasanbeig *et al.* [172] encode properties of LTL into the reward function through reward engineering using a limit deterministic Büchi automaton [173] to ensure safety. While the aforementioned methods exhibit impressive performance in ensuring safety in most tasks, their deployment in real-world applications could be challenging if environmental knowledge is unavailable. In particular, there are few studies on formal methods for safe RL since most formal methods require a dynamics model or external knowledge to define the action space.

C. Safe Multi-Agent Reinforcement Learning

Safe RL has received increasing attention both from academia and industry. However, most current RL methods are based on the single-agent setting. Safe MARL is still a relatively new area that has emerged in recent years. Little research has yet been carried out that considers the safe multi-agent RL, which can be seen as a multi-agent CMDP problem. Safe multi-agent RL not only needs to consider the ego agent's safety and reward, but also needs to take into account other agents' safety and reward. In this section, we analyze some safe multi-agent RL methods in detail. Due to the page limit, safe MARL details are provided in the preprint version of our investigations [77].

D. Summary of Safe Reinforcement Learning Methods

In this section, we introduce safe RL algorithms from different perspectives, such as safe single-agent RL, safe multi-agent RL; model-based safe RL, model-free safe RL. Even

though a number of algorithms display impressive performance in terms of reward scores, there is a long way to go toward real-world applications. In addition, based on our investigation, in MARL settings, [174] is one of the earliest methods that present convergence guarantee for multi-agent systems in year 2000. However, only a few algorithms consider safety constraints, and safe MARL still requires a lot more attention.

IV. Theory of Safe Reinforcement Learning

This section investigates theoretical techniques for analyzing safe RL. Firstly, we introduce the essential theoretical differences between standard RL and safe RL. Secondly, we investigate theories for primal-dual approaches and constrained policy optimization, two prominent methods in theoretical safe RL. Finally, we delve into safe RL sample complexity and safety violation analysis, and discuss other theoretical frameworks for safe RL. This section aims to comprehensively investigate safe RL theories by clearly delineating the theoretical foundations and analytical techniques. This exploration demonstrates the critical differences from standard RL and highlights the theoretical underpinnings of various safe RL settings and their associated analyses.

A. Difference between RL and Safe RL

The standard RL exists a deterministic stationary policy [83], while CMDP may not exist uniformly optimal stationary policies [78], which is one of the main differences between safe RL and standard RL. We present it as follows, and the blog [175] presents a two-state example to illustrate this issue.

Definition 1 (Number of Randomizations)**.** *Considering the stochastic stationary policy π, for the state s, if there exists a set $\mathcal{A}(s) \subset \mathcal{A}$ with the size $|\mathcal{A}(s)| = n + 1$, such that for each $a \in \mathcal{A}(s)$: $\pi(a|s) > 0$, we say that the total number of randomizations under state s is n, and the integer n denoted as $n_\pi(s)$ to emphasize that it depoends on the policy π and the state s. Furthermore, we say that the total number of randomizations under π is n_π, if $n_\pi = \sum_{s\in\mathcal{S}} n_\pi(s)$.*

Theorem 1. *If $\Pi_\mathcal{C} \neq \varnothing$, then for the CMDP problem, there exists an optimal policy $\pi_\star$ such that one of the following cases holds: (i) the optimal policy $\pi_\star$ is a deterministic stationary policy; (ii) or the optimal policy $\pi_\star$ is a stochastic stationary policy, and the total number of randomization of $\pi_\star$ is at most m, where m is the number of constraints.*

Let $\mathcal{A}_\star(s) := \{a : \pi_\star(a|s) > 0\}$ denote the set that collects the *optimal actions*, then according to the Definition 1 and Theorem 1, we know $|\mathcal{A}_\star(s)| \leq m + 1$.

Additionally, the next theorem shows that checking a safe policy is time-expensive.

Theorem 2. *[176] Checking feasibility in CMDPs among deterministic policies is NP-hard.*

B. Theory for Constrained Policy Optimization

This section primarily investigates representative safe RL theories grounded in constrained policy optimization due to space limitations. In future work, we plan to delve into control theory and formal methods to further enhance our understanding of safe RL theories.

A standard way to solve a CMDP problem is the Lagrangian approach [117] that is also known as primal-dual optimization,

$$(\pi_\star, \boldsymbol{\lambda}_\star) = \arg\min_{\boldsymbol{\lambda}\succeq\mathbf{0}} \max_{\pi\in\Pi_\mathrm{s}} \left\{J(\pi_{\boldsymbol{\theta}}) - \boldsymbol{\lambda}^\top(\mathbf{c}(\pi) - \mathbf{b})\right\}. \quad (5)$$

Extensive canonical algorithms are proposed to solve problem (5), e.g., [15]–[22], [177].

The work [178] presents a policy-based algorithm to solve the CMDP problem with average cost finite states, which is a stochastic approximation algorithm. Furthermore, the work [178] shows the locally optimal policy of the proposed algorithm. Chen et al. [21] propose the Reward Constrained Policy Optimization (RCPO) that is a multi-timescale algorithm for safe RL, and [21] show the asymptotical convergence of RCPO. The main idea of achieving such an asymptotical convergence is stochastic approximation [179], [180] that has been widely used in RL, e.g., [181]–[183]. For the global non-asymptotic convergence guarantees, see [112], [184]–[189]. Based on Dankin's Theorem and Convex Analysis [190], [18] provides theoretical support to the primal-dual (i.e., the Lagrange multiplier) method with a zero duality gap, which implies that the primal-dual method can be solved precisely in the dual domain.

Different from the above approach, the work CPO [3] suggests to use a surrogate cost function which evaluates the constraint $J^c(\pi_{\boldsymbol{\theta}})$ according to the samples collected from the current policy $\pi_{\boldsymbol{\theta}_k}$. Although using a surrogate function to replace the cumulative constraint cost has appeared in [166], [191], [192], CPO [3] firstly show their algorithm guarantees for near-constraint satisfaction.

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\max_{\pi_{\boldsymbol{\theta}}\in\Pi_{\boldsymbol{\theta}}} \quad \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \quad (6)$$

$$\text{s.t.} \quad J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[A^c_{\pi_{\boldsymbol{\theta}_k}}(s,a)\right] \leq b, \quad (7)$$

$$\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k}) = \mathbb{E}_{s\sim d^{\rho_0}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)}[\mathbf{KL}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})[s]] \leq \delta. \quad (8)$$

Existing recent works (e.g., [3], [23]–[27]) try to find some convex approximations to replace the terms $A_{\pi_{\boldsymbol{\theta}_k}}(s,a)$ and $\bar{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_k})$. Concretely, [3] suggest using the first-order Taylor expansion to replace (6)-(7), the second-oder approximation to replace (8). Such first-order and second-order approximations turn a non-convex problem (6)-(8) into a convex problem. This would appear to be a simple solution, but this approach results in many error sources and troubles in practice. Firstly, it still lacks a theory analysis to show the difference between the non-convex problem (6)-(8) and its convex approximation. Policy optimization is a typical non-convex problem [193]; its convex approximation may introduce some errors for its original issue. Secondly, as mentioned in the previous section, CPO updates parameters according to conjugate gradient [194], and its solution involves the inverse Fisher information matrix, which requires expensive

Model-Based Learning	Algorithm / Reference	Settings	Iteration Complexity	Safety Violation
Lower bound	[196] [197]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	/
Value-Based	OptDual-CMDP [108]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OptPrimalDual-CMDP [108]	Assumption1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	ConRL [145]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OptPess-PrimalDual [109]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	UC-CFH [51]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
Value-Based	OPDOP [49, Theorem 1]	Assumption 2,3	$\mathcal{O}\left(\frac{d\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{d\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Policy-Based	NPG-PD [53, Theorem 1]	Assumption 2	$\mathcal{O}\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$

TABLE I: This table summarizes the model-based state-of-the-art algorithms for safe RL or CMDP.

Model-Free Learning	Algorithm / Reference	Settings	Iteration Complexity	Safety Violation
Policy-Based	ConRL [145, Remark 3.5]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Value-Based	CSPDA [107][7]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Value-Based	Triple-Q [110]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^{2.5}\lvert\mathcal{A}\rvert^{2.5}}{(1-\gamma)^{18.5}\epsilon^6}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^{2.5}\lvert\mathcal{A}\rvert^{2.5}}{(1-\gamma)^{18.5}\epsilon^6}\right)$
Value-Based	Reward-Free CRL [52]	Assumption 3	$\mathcal{O}\left(\frac{\min\{d,\lvert\mathcal{S}\rvert\}\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\min\{d,\lvert\mathcal{S}\rvert\}\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^4\epsilon^2}\right)$
Policy-Based	CRPO [54, Theorem 1]	Assumption 1	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^4}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^7\epsilon^4}\right)$
Policy-Based	On-Line NPG-PD [55, Theorem 1]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^6\lvert\mathcal{A}\rvert^6}{(1-\gamma)^{12}\epsilon^6}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^6\lvert\mathcal{A}\rvert^6}{(1-\gamma)^{12}\epsilon^6}\right)$
Policy-Based	NPG-PD [53, Theorem 4]	Assumption 2	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^4\epsilon^2}\right)$	$\mathcal{O}\left(\frac{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert^2}{(1-\gamma)^4\epsilon^2}\right)$

TABLE II: This table summarizes the model-free state-of-the-art algorithms for safe RL or CMDP.

computation for each update. Later, the work [27] proposes PCPO that also uses second-order approximation, resulting in an expensive computation. In addition, the asymptotic results that safe RL methods can achieve are summarised in the preprint version of the paper [77].

C. Sample Complexity and Safety Violation

In this section, we review the sample complexity and safety violation analysis of model-based and model-free safe RL $\mathcal{O}(\epsilon)$-optimality (sample complexity and safety violation of brief introduction that we investigate studies is given in Table I and Table II), where we define a policy π of $\mathcal{O}(\epsilon)$-optimality as follows,

$$J(\pi) - J(\pi_\star) \leq \epsilon, \ \{J^c(\pi) - b\}_+ \leq \epsilon. \tag{9}$$

In this section, we review the sample complexity of the algorithms that match $\mathcal{O}(\epsilon)$-optimality.

Assumption 1 (Feasibility)**.** *The unknown CMDP is feasible, i.e., there exists an unknown policy π which satisfies the constraints. Thus, an optimal policy exists as well.*

Assumption 2 (Slater Condition, [195])**.** *There exists a vector $\xi \prec 0$, and a policy $\bar{\pi}$ such that*

$$J^c(\bar{\pi}) - b \preceq \xi. \tag{10}$$

Assumption 3 (Linear CMDP)**.** *The CMDP is a linear with a kernel feature map $\psi : \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, for each h, there exists a vector θ_h such that $\mathbb{P}(s'|s,a) = \langle \psi(s,a,s'), \theta_h \rangle$; there exists a feature map $\varphi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ and vectors $\theta_{r,h}$ and $\theta_{c,h}$, such that, $r_h(s,a) = \langle \varphi(s,a), \theta_{r,h} \rangle$ and $c_h(s,a) = \langle \varphi(s,a), \theta_{c,h} \rangle$.*

It is worth referring to [196], [197] that provide a lower bound of samples to match $\mathcal{O}(\epsilon)$ -optimality with the complexity of $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$, which is helpful for us to understand the RL safety guarantees.

1) Model-Based Safe Reinforcement Learning: Linear programming and Lagrangian approximation are widely used in model-based safe RL if the estimated transition model is either given or estimated accurately enough [198]. OptDual-CMDP [108] achieves sublinear regret with respect to the main utility while having a sublinear regret on the constraint violations, i.e., the OptDual-CMDP needs $\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ to achieve a $\mathcal{O}(\epsilon)$-optimality. the Upper-Confidence Constrained Fixed-Horizon RL method (UC-CFH) [51] provides a proximal optimal policy under the probably approximately correctness (PAC) analysis. The main idea of UC-CFH is to consider linear programming method to online learning to design an algorithm to finite-horizon CMDP. Concretely, UC-CFH [51] needs $\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ samples, and we should notice that according to [107], Theorem 1 in [51] involves a constant C that is bounded by $|\mathcal{S}|$. OptPess-PrimalDual [109] provides a way to keeps the performance with $\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ sample complexity with a known strictly safe policy. OptPess-PrimalDual [109] also claims that OptPess-PrimalDual shares a higher probability to achieve a zero constraint violation.

An Optimistic Primal-Dual proximal policy OPtimization (OPDOP) method [49] shows a bound concerning the feature mapping and the capacity of the state-action space, which leads to a sample complexity of $\mathcal{O}\left(\frac{d|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$, where d is the dimension of the feature mapping. Besides, the work [49] claims even if the dimension of state space goes to infinity, the bound also holds, which implies the merit of OPDOP. Similar techniques also be considered by [152]. An upper confidence bound value iteration (UCBVI)-γ method [50] achieves the sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\epsilon^2}\right)$ that matches the minimax lower bound up to logarithmic factors. The work [53] applies a natural policy gradient method to solve constrained Markov decision processes. The NPG-PD algorithm applies the gradient descent method to learn the primal variable, while learning the primal variable via natural policy gradient (NPG). The work [53] shows the sample complexity of NPG-PD achieves $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\epsilon^2}\right)$. We should emphasize that Theorem 1 of [53] shows a convergence rate independent on $\mathcal{S}$ and $\mathcal{A}$ due to the assumption that the agent can exactly access the policy gradient. The work [150] presents the efficient algorithms for safe RL with linear function approximation.

ConRL [145] obtains a sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\right)$. To achieves this result, the work ConRL [145] provides an analysis under two settings of strong theoretical guarantees. Firstly, [145] assumes that ConRL has a learning maximization process with the concave reward function, and this maximization falls into a convex expected value of constraints. The second setting is that during the learning maximization process, the resources never exceed specified levels. Although ConRL plays two additional settings, the complexity is still higher than previous methods, at least with a factor $\frac{1}{(1-\gamma)^2}$.

[7]The generative model is used, which needs additive samples to create a generative model.

2) Model-Free Safe Reinforcement Learning: Model-free safe RL algorithms, including IPO [142], Lyapunov-Based Safe RL [35], [36], PCPO [27], SAILR [139], SPRL [140], SNO-MDP [141], FOCOPS [120], A-CRL [168], CUP [199] and DCRL [161] all lack convergence rate analysis.

The work [53] shows NPG-PD obtains a sublinear convergence rate for both learning the reward optimality and safety constraints. NPG-PD solves the CMDP with softmax policy, where the reward objective is a non-concave and cost objective is non-convex, NPG-PD [53] shows that with a proper design, policy gradient can also obtain an algorithm that converges at a sublinear rate. Concretely, the Theorem 4 of [53] shows NPG-PD achieves the sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{(1-\gamma)^4\epsilon^2}\right)$, and we should notice that in Theorem 4 of [53], $|\mathcal{S}|^2|\mathcal{A}|^2$ samples are necessary for the two outer loops. Later, the work [55] extends the critical idea of NPG-PD and proposes an online version of NPG-PD that needs the sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}|^6|\mathcal{A}|^6}{(1-\gamma)^{12}\epsilon^6}\right)$, where we show the iteration complexity after some simple algebra according to [55, Lemma 8-9]. Clearly, online learning NPG-PD [55] needs additional $\mathcal{O}(\epsilon^{-4})$ trajectories than NPG-PD [55].

The work [54] proposed a primal-type algorithmic framework to solve SRL problems, and they show the proposed algorithm needs $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^7\epsilon^4}\right)$ sample complexity to obtain $\mathcal{O}(\epsilon)$-optimality, where we notice that the inner loop with $K_{\text{in}} = \mathcal{O}\left(\frac{T}{(1-\gamma)|\mathcal{S}||\mathcal{A}|}\right)$ iteration is needed [54, Theorem 3].

The work [107] proposes the CSPDA algorithm needs the sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$. However, the inner loop of their Algorithm 1 needs an additional generative model. Triple-Q [110] needs the sample complexity of $\mathcal{O}\left(\frac{|\mathcal{S}|^{2.5}|\mathcal{A}|^{2.5}}{(1-\gamma)^{18.5}\epsilon^5}\right)$. We show this iteration complexity according to a recent work [107]. Since the work [110] study the finite-horizon CMDP, we believe their Triple-Q plays at least $\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|^2}{\epsilon^5}\right)$, which is still higher than NPG-PD [53] at least with a factor $\mathcal{O}(\epsilon^{-3})$. The work [52] proposes a safe RL algorithm that needs $\mathcal{O}\left(\frac{\min\{d,|\mathcal{S}|\}|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$. It is noteworthy that we show the sample complexity here for the worst-case of constraint violation shown in [52] reaches $\mathcal{O}\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^4\epsilon^2}\right)$, if the dimension of the feature mapping d is greater than $|\mathcal{S}|$. The work [200] provides an analysis of model-free regret-optimal best policy identification for online CMDPs.

V. Applications of Safe Reinforcement Learning

RL applications for challenging tasks have a long tradition. Some RL methods are leveraged to solve complex problems before neural network learning arises. For example, TD learning is used to solve problems including backgammon playing [201], [202], job-shop scheduling and elevator group control [203]; a stochastic approximation algorithm with RL properties is utilized to solve pricing options for financial derivatives in two-player and zero-sum games [204]. However, most of the above methods are on a small scale or have linear settings, and most of the problems they solve are discrete. The policy values are almost approximated to address more challenging tasks for large-scale, continuous, and high-dimensional problems, e.g., using neural networks is currently a widely adopted method to learn sophisticated policy strategies in modern RL. In this section, to investigate the **"Safety Applications"** problem, we introduce safe RL applications, including tabular-setting RL, and modern RL applications, such as autonomous driving, robotics, and recommendation systems.

The application methods share high-level techniques with Section III: methods of safe RL. However, application methods are more focused on specific applications and how the methods are deployed in these domains, such as autonomous driving and robotics. For instance, in autonomous driving applications, greater emphasis is placed on the application environment and how to ensure planning safety for autonomous vehicles with safe RL. In contrast, the safe RL method section primarily discusses technical solutions to ensure learning safety, and these methods can be deployed in various applications, e.g., autonomous driving and robotics.

The distinction between the two sections lies in their scope and emphasis. While Section III explores the theoretical and algorithmic underpinnings of safe RL methods, the application methods section delves into the practical considerations and challenges associated with implementing these techniques in real-world scenarios. This section aims to bridge the gap between theoretical frameworks and their practical implementation, investigating the nuances and domain-specific requirements of different application domains. By introducing the deployment of safe RL methods in specific applications, this section provides valuable insights into the real-world constraints, environmental factors, and safety-critical considerations that must be addressed.

A. Safe Reinforcement Learning for Autonomous Driving

More recently, many methods have been proposed for autonomous driving based on modern, advanced techniques. The work [205], proposed by Pomerleau, may be one of the first learning-based methods for autonomous driving, developed in 1989. Gu *et al.* [57] provide a motion planning method for automated driving based on constrained RL. They combine traditional motion planning and RL methods to perform better than pure RL or traditional methods. Specifically, the topological path search [206], [207] and trajectory lane model, which is derived from trajectory units [208]–[210], are leveraged to constrain the RL search space. Their method can be used very well for corridor scenarios that consider environmental uncertainty.

In contrast to Gu *et al.* [57], Wen *et al.* [211] provide a parallel safe RL method for vehicle lane-keeping and multi-vehicle decision-making tasks by using pure constrained RL methods. They extend an actor-critic framework to a three-layer-neural-network framework by adding a risk layer for autonomous driving safety. The synchronized strategy is used to optimize parallel policies for better searching viable states and speeding up convergence.

Krasowski *et al.* [59] develop a safe RL framework for autonomous driving motion planning, in which they focus more on the high-level decision-making problems for lane changes of vehicles on highways. Based on the work [59], Wang [64] presents a low-level decision-making method via a safety layer of CBFs [212], and a legal safe control method

by following traffic rules to ensure motion planning safety for autonomous driving in highway scenarios. Different from Wang's method [64] using CBFs, Cao *et al.* [213] improve the safety of autonomous driving in low-level decision-making settings by integrating a rule-based policy, e.g., a Gipps car-following model [214], into RL framework, and a Monte Carlo tree searching method [215] is used to generate their RL framework policies. Although safe RL has achieved considerable success in low-level decision-making, it still falls short in guaranteeing autonomous driving safety within complex environments, particularly when encountering multiple dynamic and uncertain obstacles.

Furthermore, Mirchevsk *et al.* [62] leverage a Q learning method [216] and a tree-based ensemble method [217] used as a function approximator, to achieve high-level control for lane changes in highway scenarios. Their method has shown impressive performance by reducing collision probability. Nevertheless, this method may only be suitable for two-lane changing environments, since one-lane change options are only considered in the environments at any time. Furthermore, Mirchevsk *et al.* [63] use formal methods [218] to guarantee safety when they use RL for the safe and high-level planning of autonomous driving in autonomous lane changes. Therefore, their method can be used for more complex environments compared to the work [62], and their method displays good performance in highway scenarios with an arbitrary number of lanes. They also integrate safety verification into RL methods to guarantee agent action safety.

In [219], similar to Mirchevsk *et al.* [63], they introduce a verification method for RL safety, more particularly, they verify the action safety. The policy can be learned adaptively in a distributional RL framework. To ensure action safety, Isele *et al.* [58] use prediction methods to render safe RL exploration for intersection behaviors during autonomous driving. Remarkably, they can constrain agents' actions by prediction methods in multi-agent scenarios, where they assume other agents are not adversarial and an agent's actions are generated by a distribution. Kendall *et al.* [61] provides a model-free based RL method, which is combined by variational autoencoder [220] and DDPG [221]. Their method may be one of the first to implement real-world vehicle experiments using RL, in which they use logical rules to achieve autonomous driving safety, and use mapping and direct supervision to navigate the vehicles.

Moreover, Kalweit *et al.* [60] develop an off-policy and constrained Q learning method for high-level autonomous driving in simulation environments. They use the transportation software, SUMO [222], as a simulation platform and the real high dimensional data set [223] to verify the effectiveness of their methods. Specifically, they constrain the agent's action space when the agent performs a Q value update; the safe policy is then searched for autonomous driving. Different from the above perspectives, Atakishiyev *et al.* introduce some Explainable Artificial Intelligence (XAI) methods, and a framework for safe autonomous driving [224], in which Explainable Reinforcement Learning (XRL) for choosing vehicle actions is mentioned. Although XRL can be helpful in promoting the development of safe and trustworthy autonomous systems, this topic has just been studied with regard to safe RL, and the relevant research is not remarkably mature.

B. Safe Reinforcement Learning for Robotics

Some learning methods for robot applications have shown excellent results [225], [226]. However, the methods do not explicitly consider the agent's safety as an optimization objective. Particularly, there are a number of works that apply RL methods to simulation robots or real-world robots. Nonetheless, most of them do not take safety into account during the learning process. For the purpose of better applications using RL methods, we need to figure out how to design safe RL algorithms to achieve better performance for real-world applications. Safe RL is a bridge that tries to improve the safe learning efficiency and connects the RL simulation experiments to real-world applications in robotics.

For example, in [69], Slack *et al.* use an offline primitive learning method, called SAFER, to improve safe RL data efficiency and safety rate. However, SAFER has not theoretical safety guarantees. They collected safe trajectories as a safe set by a scripted policy [227], and applied the safe trajectories to a learning process. In terms of safety and success rate, their method has achieved better performance in PyBullet [228] simulation experiments than other baselines.

To facilitate the deployment of safe RL in real-world scenarios, beyond merely simulation-based experiments, the fully differentiable OptLayer [68] is developed to ensure safe actions that the robots can only take. More importantly, they implement their methods in real-world robots using a 6-DoF industrial manipulator and have received significant attention. However, the method may be limited in high-dimensional space for robot manipulations since the OptLayer may not be able to optimize policies efficiently in complex tasks, especially in high-dimensional space. Furthermore, Garcia and Fernandez [65] present a safe RL algorithm based on safe exploration, in which they develop a smoother and continuous risk function for safe exploration. It can guarantee a monotonic increase, and the risk function is used to help follow a baseline policy. Building on this framework, subsequent work [66] has successfully adapted this risk function-based algorithm for practical applications in real-world robotics.

In contrast to safety layer methods or policy optimization, a series of safe RL methods grounded in control theory have been proposed for robot learning. Notably, Perkins and Barto [67] employ Lyapunov functions, which are developed over a century ago [229], traditionally used to ensure the stability of controllers [230]. They specifically utilize a Lyapunov function to constrain the action space, effectively ensuring the safety of all policies and maintaining agent performance. Moreover, they formulate a set of control laws with predefined Lyapunov domain knowledge. Their approach, applying a Lyapunov-based safe RL method to pendulum tasks using a robot arm in simulated environments, may be among the initial applications of Lyapunov functions for safe RL in robotics. Although Lyapunov functions can enhance system safety and stability, a significant challenge remains: designing a function that fulfills all policy safety requirements necessitates a deep understanding of the system dynamics and the specific domain knowledge of Lyapunov functions.

Similarly, several safe RL methods that require system models are developed in recent years. In particular, Thomas *et al.* [70] leverage a model-based RL to achieve the agents' safety by incorporating the model knowledge. Specifically, they use the agents' dynamics to anticipate the trajectories of the next few steps, and thus prevent agents from entering unsafe states or performing unsafe actions. Based on their method, they apply the proposed method for MuJoCo robot control in simulation environments. Their method may be more suitable for short-horizon trajectories. However, if it encounters a large-scale horizon, the method may not work well since it needs to plan the next few steps quickly. Furthermore, in [231], Liu *et al.* provide a safe exploration method for robot learning on the constrained manifold. More specifically, the robot models and manifold constraints in tangent space are utilized to help ensure robot safety during the RL exploration process. Their method can leverage any model-free RL methods for robot learning on the constrained manifold, since the constrained problem is converted as an unconstrained problem in tangent space, and their method can search for policy on the exploration of safe regions. Nonetheless, an accurate robot model and tracking controller are required in their method, which may not be suitable for real-world applications.

C. Safe Reinforcement Learning for Other Applications

Apart from autonomous driving and robotics of safe RL applications, safe RL is also adopted to ensure safety in recommender systems [232], video compression [71], video transmissions [233], wireless security [234], satellite docking [235], edge computing [236], chemical processes [44] and vehicle schedule [1], [72], and so on. In recommender systems, for example, Singh *et al.* [232] deploy safe RL to optimize the healthy recommendation sequences of recommender systems by utilizing a policy gradient method algorithm on the Conditional Value at Risk (CVaR) method [237], whereby they optimize positive feedback while constraining cumulative exposure of health risk. In video compression, Mandhane *et al.* [71] leverage the Muzero [238], one of the alpha series algorithms, to solve the safe RL problem in video compression. More specifically, as shown in function (11), they optimize the encoded video quality by maximizing quantization parameters (QPs) via policy learning while satisfying the bit rate constraints. Their experiments proved that their method can achieve better performance than traditional methods and related modern machine learning methods on the YouTube UGC dataset [239]. However, the method may not be easily scalable for large-scale datasets.

$$\max_{\text{QPs}} \text{Encoded Video Quality} \quad \text{s.t. Bitrate} \leq \text{Target} \tag{11}$$

In wireless security [234], based on the Inter-agent transfer learning method [240], Lu *et al.* develop a safe RL method for wireless security using a hierarchical structure. To be more specific, the target Q network and E-networks with CNN [241] are used to optimize the stability of policy exploration, and reduce the risk of the policy exploration, ultimately enhancing the wireless security in UAV communication against jamming.

D. Summary of Applications

In this section, we analyze safe RL methods for autonomous driving and robotics and so on, whereby guaranteeing safety and improving reward simultaneously during model training is a challenging problem. Some methods are proposed to deal with this problem, such as model-based safe RL to plan safe actions [70], Lyapunov function to guarantee the safety of agents [67], predefined baseline policy for safe exploration [65], formal verification for safe autonomous driving [63], constrained Q learning for high-level vehicle lane changes [60], etc. Although these methods have been very successful, one major problem that remains is how to rigorously guarantee safety during exploration and retain the reward of monotonic improvement, and how to guarantee stability and convergence when safe RL methods are applied to real-world applications. (Due to page limit, the detail of robot applications is provided in the preprint version of the paper [77].)

VI. Benchmarks of Safe Reinforcement Learning

Several safety benchmarks for safe RL have been developed, and various baselines have been compared on the safe RL benchmarks [74]. More importantly, the safe RL benchmarks, including single-agent and multi-agent benchmarks, have made massive contributions to the RL community and helped safe RL move toward real-world applications. In this section, we investigate the popular safe RL benchmarks and try to answer the **"Safety Benchmarks"** problem.

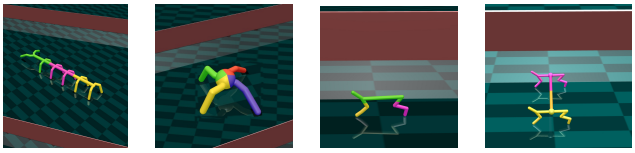
A. Benchmarks of Safe Single-Agent Reinforcement Learning

1) AI Safety Gridworlds: AI Safety Gridworlds [73] [8] is a kind of 2-D environment that is used to evaluate safe RL algorithms. All of the environments are based on the 10X10 grids. An agent is arranged in one cell of the grid, and obstacles are arranged in some cells. The action space is discrete in AI safety Gridworlds. An agent can take action from action space $A = \{right, left, up, down\}$.

2) Safety Gym: Safety Gym [74] [9] is based on Open AI Gym [242] and MuJoCo [243] environments. It also takes into account 2-D environments in different tasks, e.g., a 2-D robot such as a Point robot, a Car robot, or a Doggo robot can turn and move to navigate a goal position while avoiding crashing into unsafe areas in a 2-D plane. Moreover, the robot's actions are continuous. There are many types of costs associated with the safety gym. For example, the robot has to avoid crashing into dangerous areas, non-goal objects, immobile obstacles, and moving objects. Otherwise, costs will be incurred.

3) Safe Control Gym: Aiming at safe control and learning problems, Yuan *et al.* propose safe control gym [244] [10], an extension benchmark of OpenAI Gym [242], which integrates traditional control methods [245], learning based-control methods [246] and RL methods [247], [248] into a framework, in which model-based and data-based control approaches are both supported. They mainly consider the cart-pole task, 1D and 2D

[8] https://github.com/deepmind/ai-safety-gridworlds.git
[9] https://github.com/openai/safety-gym.git
[10] https://github.com/utiasDSL/safe-control-gym.git



(a) (b) (c) (d)

Fig. 2: Example tasks in Safe Multi-Agent MuJoCo Environments, such as six-agent ManyAgent Ant (a), four-agent Ant (b), three-agent HalfCheetah (c), and two-agent HalfCheetah (d) tasks. Body parts of different colours are controlled by different agents. Agents jointly learn to manipulate the robot, while avoiding crashing into unsafe red areas, for details, see [75] (Adapt the figures with permission from [75]).

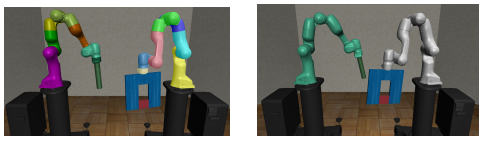
quadrotor tasks, stabilization, and trajectory tracking tasks in their environments. Compared with Safety Gym [74] and AI safety gridworlds [73], safe control gym [244] may be more suitable for sim-to-real research, since they offer numerous options for implementing non-idealities that resemble real-world robotics, such as randomization, dynamics disturbances and also support a symbolic framework to present systems' dynamics and constraints.

B. Benchmarks of Safe Multi-Agent Reinforcement Learning

In recent years, three safe multi-agent benchmarks have been developed by [75], namely Safe Multi-Agent MuJoCo (Safe MAMuJoCo), Safe Multi-Agent Robosuite (Safe MARobosuite), Safe Multi-Agent Isaac Gym (Safe MAIG), respectively. The safe multi-agent benchmarks can help promote the research of safe MARL.

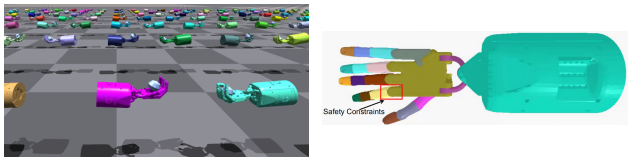
1) Safe Multi-Agent MuJoCo: Safe MAMuJoCo [75] [11] is an extension of MAMuJoCo [249]. In Safe MAMuJoCo, safety-aware agents have to learn not only the skillful manipulations of a robot, but also how to avoid crashing into unsafe obstacles and positions. In particular, the background environment, agents, physics simulator, and reward function are preserved. However, unlike its predecessor, a Safe MAMuJoCo environment comes with obstacles like walls or bombs. Furthermore, with the increasing risk of an agent stumbling upon an obstacle, the environment emits cost [242]. According to the scheme in [250], the cost functions are characterized for each task; the examples of Safe MAMuJoCo robots are shown in Figure 2.

2) Safe Multi-Agent Robosuite: Safe Multi-Agent Robotsuite (Safe MARobosuite) [75] [12], shown in Figure 3, has been developed on the basis of Robosuite [251] which is a popular robotic arm benchmark for single-agent reinforcement learning. In Safe MARobosuite, multiple agents are set up according to the robot joint settings, and each agent controls every joint or several joints. A Lift task, for example, can be divided up into 2 agents (2x4 Lift), 4 agents (4x2 Lift), 8 agents (8x1 Lift). More importantly, Safe MARobosuite can be easily used for modular robots and make robots have good robustness and scalability. For instance, when communication bandwidth is limited, or some joints of robotic arms are broken, causing malfunctioning communication, modular robots can still work well. The reward setting is the same as Robosuite [251], and the cost design is used to prevent robots from crashing into unsafe areas.



(a) (b)

Fig. 3: Example tasks in Safe MARobosuite Environments, e.g., fourteen-agent Lift tasks (a), and two-agent Lift tasks (b). Agents jointly learn to manipulate the robot, while avoiding crashing into unsafe areas, for details, see [75](Adapt the figures with permission from [75]).



(a) (b)

Fig. 4: Safe multi-agent Isaac Gym environments. Different robot body parts of different colours are manipulated by different agents in environments. Agents jointly learn how to control the robot (a), whilst the safety constraints (b) are not violated (Adapt the figures with permission from [75]).

3) Safe Multi-Agent Isaac Gym: Safe Multi-Agent Isaac Gym (Safe MAIG) [13] is a high-performance environment that uses GPUs for trajectory sampling and logical computation based on Isaac Gym [252], an example task is shown in Figure 4. The computation speed of Safe MAIG is almost ten times that of Safe MAuJoCo and Safe MARobosuite on the same settings. The communication speed is also better than Safe MAMuJoCo and Safe MARobosuite. However, the memory might need to be optimized between CPUs and GPUs to achieve better sample efficiency.

VII. Challenges and Outlook

When deploying RL in real-world applications, numerous challenges emerge throughout the implementation process. In this section, the **"Safety Challenges"** problem is investigated by proposing several significant challenges we need to address moving forward. Additionally, we identify potential research directions for enhancing the safety of RL systems.

A. Human-Compatible Safe Reinforcement Learning

Modern safe RL algorithms rely on humans to state the objectives of safety. While humans understand the dangers, the potential risks are less noticed. Below we discuss two challenges in human preference statements and concerns on ethics and morality.

Human preference statement. Many challenges are posed as AI agents are frequently evaluated in terms of performance measures, such as human-stated rewards. On the one hand, while it is usually assumed that humans are acting honestly in

[11] https://github.com/chauncygu/Safe-Multi-Agent-Mujoco.git
[12] https://github.com/chauncygu/Safe-Multi-Agent-Robosuite.git
[13] https://github.com/chauncygu/Safe-Multi-Agent-Isaac-Gym.git

specifying their preferences, such as by rewards or demonstrations, the consequence of humans misstating their objectives is commonly underestimated. Humans may maliciously or unintentionally misstate their preferences, leading the safe RL agent to perform unexpected implementations. One example is the Tay chatbot from Microsoft; prankster users falsify their demonstrations and train Tay to mix racist comments into its dialogue [253]. On the other hand, multiple humans might be involved in training one safe RL agent. Thus, agents have to learn to strike a balance between the widely different human preferences. Earlier attempt [6] considers one agent vs one human scenario. However, many open questions remain, such as training robust agents against malicious users, personalizing assistance toward human preferences, etc.

Ethics and morality concerns. In modern society, the human interrelationship is built based on social or moral norms. While reinforcement learning agents are deployed to the real world, they start having impacts on each other, turning into a multi-agent system, in which norms act similarly in human society on agents. Therefore, the decisions made by agents always involve ethical issues. For example, social dilemmas will emerge from the relation between individual goals and overall interests [254], [255]. The conflicts are produced when each agent aims to maximize its benefit. For another example, consider a *trolley problem* [256]. When the agent is faced with the choice of either harming multiple people on the current track or one person by diverting the train, what would you expect the safe agents to choose? Or, more realistically, when the driving agent is about to bump into a lorry, it can swerve off the road to the left to save itself. However, there is a bike on the left. How could the driving agent make decisions? A human driver's knee-jerk reaction might be swerving left to save itself. However, the decision of the driving agent depends on its value systems. How to leverage the different value systems to enable safe agents to make ethical decisions is an open question.

B. Industrial Deployment Standards for Safe Reinforcement Learning

Although safe RL has been developed with a wealth of well-understood methods and algorithms in recent years [85], [86], [257], to our knowledge, there is no RL deployment standard for industrial applications, including technical standards and morality standards, etc. More attention is required on deployment standards and the alignment between academia and industries. Applications include robotics, autonomous driving, recommendation systems, finance, *et al*. We should tailor specific deployment standards to specific applications.

Technical standards. In a technical standard, we need to think about how much efficiency RL can generate, how much time and money can be saved using RL methods, what environments can be handled with RL, how to design cost and reward functions considering the balance between RL reward, performance and safety, etc.

Law standards. Human-machine interaction needs to be considered in legal judgments. For example, when robots hurt humans due to programming errors using RL methods, we need to determine how responsibilities are divided, e.g., do programmers of robots need to take more responsibility, or do robot users need to take more responsibility?

C. Safety Guarantee Efficiently in Large Number of Agents' Environments

Since the decision-making space is incredibly large when the number of agents increases in a safe MARL setting, it is not easy to optimize the multi-agent policy to finish tasks safely [258]. Thus, efficiently guaranteeing safety in an environment with a large number of agents, e.g., 1 million agents, is still a challenging problem.

Theory of safe MARL. Theory and experiments of safe MARL for massive swarm robots should be considered in the future, e.g., the convergence problem remains open for massive safe MARL methods without strong assumptions. Furthermore, optimizing sample complexity and stability within safe MARL environments is essential. Specifically, greater emphasis should be placed on the following key points regarding the theory of safe MARL. (1). Credit assignment both in cost and reward. In cooperative, competitive, and mixed game settings, it is crucial to contemplate the trade-off between reward and safety performance, e.g., a policy should be searched to minimize each agent's cost value while improving reward. Furthermore, it is necessary to optimize the precise cost value for each agent, and consider each agent's cost credit. (2). Nonstability. In a multi-agent system, when an agent takes actions, which will influence other agents' actions and may make other agents get worse reward value or unsafe. (3). Scalability. In safe MARL settings, when the number of agents becomes large, such as one billion agents, it will be challenging for computation and hard to ensure agents' safety, since it is almost impossible to estimate each agent's Q values or V value simultaneously.

Multiple heterogeneous constraints. It still needs to be determined how multiple heterogeneous constraints should be handled in multi-agent systems for safe MARL. To our knowledge, almost no study has investigated multiple heterogeneous constraints for MARL. For example, when different agents encounter different heterogeneous constraints, we need to study how to balance different heterogeneous constraints to optimize different agents' policies for safety while improving reward values.

Carry out complex multi-agent tasks safely. For instance, swarm robots perform the encirclement task, and then rescue the hostages from the dangerous scene safely. Currently, most safe MARL algorithms could have some challenging issues while conducting complex tasks, e.g., time-consuming issues or convergence issues.

Robust MARL. One of the concerns in robust MARL settings: ensuring zero cost in different tasks without tuning parameters using the same algorithms is still open [56]. Another concern is when we transfer the safe MARL simulation results to real-world applications, it is still unclear regarding how to handle the mismatch between the nominal and real system, since there may be the sensor noise [219] generated by fault sensors or disturbing sensor information transferred by adversary attacks.

Trade-off Balances. The trade-off balance between exploration and exploitation is a dilemma in RL or MARL. Safe RL

or safe MARL has the same problem. More particularly, there is another dilemma that is the trade-off between reward and cost, which is different from the exploration and exploitation, since each action can result in the change of reward and cost simultaneously, it is a multi-objective problem. Thus, in safe MARL settings, addressing the two balances mentioned earlier is imperative. In competitive game settings, efficiently modeling an opponent's decisions with limited information and implementing safe actions must be ascertained. In cooperative game settings, searching for a policy to guarantee the whole team's reward monotonic improvement while adhering to individual agent constraints is essential. In mixed-game settings, the optimization of both local and overall rewards while executing safe actions needs to be determined.

D. Possible Directions for Future Safe Reinforcement Learning Research

1) Safe MARL with game theory. Addressing the aforementioned challenges by leveraging game theory constitutes a primary approach, as various games can be considered for real-world applications across different settings, including cooperative and competitive scenarios. Optimizing safety within extensive form games also proves beneficial for practical applications. For example, in a fencing competition, it is crucial to determine methods for ensuring that both agents successfully accomplish their objectives while maintaining safety throughout the gameplay.

2) Safe RL with information theory. Information theory can be instrumental in handling uncertain reward signals and cost estimation, as well as effectively addressing challenges in large-scale MARL environments. For instance, leveraging information coding theory allows for the construction of representations for various agent actions or reward signals, thus enhancing overall efficiency.

3) Other potential directions include safeMARL inspired by human brain theory and biological insights, learning safe and diverse behaviors from human feedback (similar to ChatGPT[14]), and human-robot interactions. These directions are discussed in detail in the preprint version of our investigations [77].

VIII. Conclusion

We carefully review safe RL methods from the past 20 years, attempt to answer the key safe RL question around the investigation of safety research with **"2H3W"** problems, and provide a clear clue for further safe RL research. Firstly, five critical safe RL problems are posed, and the model-based and model-free RL methods are analyzed in a unified safety framework. Secondly, the sample complexity and convergence of each method are investigated briefly. Thirdly, applications of safe RL are analyzed, for example, in the fields of autonomous driving and robotics. Fourthly, the benchmarks for safe RL communities are revealed, which may help RL take further toward real-world applications. Finally, the challenging problems that confront us during RL applications in safe RL domains are pointed out for future research.

[14]https://openai.com/blog/chatgpt/

References

[1] R. Basso, B. Kulcsár, I. Sanchez-Diaz, and X. Qu, "Dynamic stochastic electric vehicle routing with safe reinforcement learning," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 157, p. 102496, 2022.

[2] X. Zhao, C. Gu, H. Zhang, X. Yang, X. Liu, H. Liu, and J. Tang, "Dear: Deep reinforcement learning for online advertising impression in recommender systems," in *Proc. the AAAI Conf. Artif. Intell.*, vol. 35, no. 1, 2021, pp. 750–758.

[3] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 70, 2017, pp. 22–31.

[4] A. Stevenson, *Oxford dictionary of English*. Oxford University Press, USA, 2010.

[5] A. Hans, D. Schneegaß, A. M. Schäfer, and S. Udluft, "Safe exploration for reinforcement learning." in *ESANN*. Citeseer, 2008, pp. 143–148.

[6] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," *Adv. Neur. In.*, vol. 29, 2016.

[7] G. Irving, P. Christiano, and D. Amodei, "Ai safety via debate," *arXiv:1805.00899*, 2018.

[8] J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini, and S. Legg, "Scalable agent alignment via reward modeling: a research direction," *arXiv:1811.07871*, 2018.

[9] T. M. Moldovan and P. Abbeel, "Safe exploration in markov decision processes," in *Proc. the 29th International Coference on Int. Conf. Mach. Learn.*, 2012, pp. 1451–1458.

[10] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv:1606.06565*, 2016.

[11] S. Mannor and J. N. Tsitsiklis, "Mean-variance optimization in markov decision processes," in *Proc. the 28th Int. Conf. Mach. Learn.*, 2011, pp. 177–184.

[12] E. Delage and S. Mannor, "Percentile optimization in uncertain markov decision processes with application to efficient exploration," in *Proc. the 24th Int. Conf. Mach. Learn.*, 2007, pp. 225–232.

[13] T. M. Moldovan and P. Abbeel, "Risk aversion in markov decision processes via near optimal chernoff bounds." in *NIPS*, 2012, pp. 3140–3148.

[14] G. Dulac-Arnold, D. Mankowitz, and T. Hester, "Challenges of real-world reinforcement learning," *arXiv:1904.12901*, 2019.

[15] Y. Chen, J. Dong, and Z. Wang, "A primal-dual approach to constrained markov decision processes," *arXiv:2101.10895*, 2021.

[16] H. Le, C. Voloshin, and Y. Yue, "Batch policy learning under constraints," in *Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 3703–3712.

[17] Q. Liang, F. Que, and E. Modiano, "Accelerated primal-dual policy optimization for safe reinforcement learning," *arXiv:1802.06480*, 2018.

[18] S. Paternain, L. F. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2019.

[19] R. H. Russel, M. Benosman, and J. Van Baar, "Robust constrained-mdps: Soft-constrained robust policy optimization under model uncertainty," *arXiv:2010.04870*, 2020.

[20] H. Satija, P. Amortila, and J. Pineau, "Constrained markov decision processes via backward value functions," in *Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 8502–8511.

[21] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *Int. Conf. Learn. Represent. (ICLR)*, 2019.

[22] T. Xu, Y. Liang, and G. Lan, "A primal approach to constrained policy optimization: Global optimality and finite-time analysis," *arXiv:2011.05869*, 2020.

[23] H. Bharadhwaj, A. Kumar, N. Rhinehart, S. Levine, F. Shkurti, and A. Garg, "Conservative safety critics for exploration," in *Int. Conf. Learn. Represent. (ICLR)*, 2021.

[24] L. Bisi, L. Sabbioni, E. Vittori, M. Papini, and M. Restelli, "Risk-averse trust region optimization for reward-volatility reduction," in *Proc. the Twenty-Ninth Int. Jt. Conf. Artif. Intell., IJCAI-20*, C. Bessiere, Ed., 2020, pp. 4583–4589.

[25] M. Han, L. Tian, Yuanand Zhang, J. Wang, and W. Pan, "Reinforcement learning control of constrained dynamic systems with uniformly ultimate boundedness stability guarantee," *arXiv:2011.06882*, 2020.

[26] Q. Vuong, Y. Zhang, and K. W. Ross, "Supervised policy update for deep reinforcement learning," in *Int. Conf. Learn. Represent. (ICLR)*, 2019.

[27] T.-Y. Yang, J. Rosca, K. Narasimhan, and P. J. Ramadge, "Projection-based constrained policy optimization," in *Int. Conf. Learn. Represent. (ICLR)*, 2020.

[28] G. Anderson, A. Verma, I. Dillig, and S. Chaudhuri, "Neurosymbolic reinforcement learning with formally verified exploration," *Adv. Neur. In.*, vol. 33, pp. 6172–6183, 2020.

[29] J. J. Beard and A. Baheri, "Safety verification of autonomous systems: A multi-fidelity reinforcement learning approach," *arXiv:2203.03451*, 2022.

[30] N. Fulton and A. Platzer, "Safe reinforcement learning via formal methods: Toward safe control through proof and learning," in *Proc. the AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[31] N. Hunt, N. Fulton, S. Magliacane, T. N. Hoang, S. Das, and A. Solar-Lezama, "Verifiably safe exploration for end-to-end reinforcement learning," in *Proc. the 24th International Conference on Hybrid Systems: Computation and Control*, 2021, pp. 1–11.

[32] J. Riley, R. Calinescu, C. Paterson, D. Kudenko, and A. Banks, "Reinforcement learning with quantitative verification for assured multi-agent policies," in *13th International Conference on Agents and Artificial Intelligence*. York, 2021.

[33] H. Zhu, Z. Xiong, S. Magill, and S. Jagannathan, "An inductive synthesis framework for verifiable reinforcement learning," in *Proc. the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2019, pp. 686–701.

[34] J. Choi, F. Castañeda, C. J. Tomlin, and K. Sreenath, "Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions," in *Robot. Sci. Syst. (RSS)*, 2020.

[35] Y. Chow, O. Nachum, E. Duenez-Guzman, and M. Ghavamzadeh, "A lyapunov-based approach to safe reinforcement learning," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2018.

[36] Y. Chow, O. Nachum, A. Faust, E. Duenez-Guzman, and M. Ghavamzadeh, "Lyapunov-based safe policy optimization for continuous control," *ICML Workshop RL4RealLife*, 2019.

[37] S. Huh and I. Yang, "Safe reinforcement learning for probabilistic reachability and safety specifications: A lyapunov-based approach," *arXiv:2002.10126*, 2020.

[38] A. B. Jeddi, N. L. Dehghani, and A. Shafieezadeh, "Lyapunov-based uncertainty-aware safe reinforcement learning," *arXiv:2107.13944*, 2021.

[39] A. K. Akametalu, J. F. Fisac, J. H. Gillula, S. Kaynama, M. N. Zeilinger, and C. J. Tomlin, "Reachability-based safe learning with gaussian processes," in *53rd Proc. IEEE Conf. Decis.* IEEE, 2014, pp. 1424–1431.

[40] M. Cai and C.-I. Vasile, "Safe-critical modular deep reinforcement learning with temporal logic through gaussian processes and control barrier functions," *arXiv:2109.02791*, 2021.

[41] A. I. Cowen-Rivers, D. Palenicek, V. Moens, M. A. Abdullah, A. Sootla, J. Wang, and H. Bou-Ammar, "Samba: Safe model-based & active reinforcement learning," *Mach. Learn.*, pp. 1–31, 2022.

[42] J. Fan and W. Li, "Safety-guided deep reinforcement learning via online gaussian process estimation," *arXiv:1903.02526*, 2019.

[43] K. Polymenakos, A. Abate, and S. Roberts, "Safe policy search using gaussian process models," in *Proc. the 18th Int. Conf. AAMAS*, 2019, pp. 1565–1573.

[44] T. Savage, D. Zhang, M. Mowbray, and E. A. D. R. Chanona, "Model-free safe reinforcement learning for chemical processes using gaussian processes," *IFAC-PapersOnLine*, vol. 54, no. 3, pp. 504–509, 2021.

[45] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in *Int. Conf. Mach. Learn.* PMLR, 2015, pp. 997–1005.

[46] M. Turchetta, F. Berkenkamp, and A. Krause, "Safe exploration in finite markov decision processes with gaussian processes," *Adv. Neur. In.*, vol. 29, 2016.

[47] A. Wachi, Y. Sui, Y. Yue, and M. Ono, "Safe exploration and optimization of constrained mdps using gaussian processes," in *Proc. the AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

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

[49] D. Ding, X. Wei, Z. Yang, Z. Wang, and M. Jovanovic, "Provably efficient safe exploration via primal-dual policy optimization," in *Int. Conf. Artif. Intell. Stat.* PMLR, 2021, pp. 3304–3312.

[50] J. He, D. Zhou, and Q. Gu, "Nearly minimax optimal reinforcement learning for discounted mdps," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2021.

[51] K. C. Kalagarla, R. Jain, and P. Nuzzo, "A sample-efficient algorithm for episodic finite-horizon mdp with constraints," *Proc. the AAAI Conf. Artif. Intell.*, vol. 35, no. 9, pp. 8030–8037, May 2021.

[52] S. Miryoosefi and C. Jin, "A simple reward-free approach to constrained reinforcement learning," *arXiv:2107.05216*, 2021.

[53] D. Ding, K. Zhang, T. Basar, and M. R. Jovanovic, "Natural policy gradient primal-dual method for constrained markov decision processes." in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2020.

[54] T. Xu, Y. Liang, and G. Lan, "Crpo: A new approach for safe reinforcement learning with convergence guarantee," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 11 480–11 491.

[55] S. Zeng, T. T. Doan, and J. Romberg, "Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained markov decision processes," *arXiv:2110.11383*, 2021.

[56] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, "Unsolved problems in ml safety," *arXiv:2109.13916*, 2021.

[57] S. Gu, G. Chen, L. Zhang, J. Hou, Y. Hu, and A. Knoll, "Constrained reinforcement learning for vehicle motion planning with topological reachability analysis," *Robotics*, vol. 11, no. 4, 2022.

[58] D. Isele, A. Nakhaei, and K. Fujimura, "Safe reinforcement learning on autonomous vehicles," in *2018 IEEE Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2018, pp. 1–6.

[59] H. Krasowski, X. Wang, and M. Althoff, "Safe reinforcement learning for autonomous lane changing using set-based prediction," in *2020 IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*. IEEE, 2020, pp. 1–7.

[60] G. Kalweit, M. Huegle, M. Werling, and J. Boedecker, "Deep constrained q-learning," *arXiv:2003.09398*, 2020.

[61] A. Kendall, J. Hawke, D. Janz, P. Mazur, D. Reda, J.-M. Allen, V.-D. Lam, A. Bewley, and A. Shah, "Learning to drive in a day," in *2019 Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2019, pp. 8248–8254.

[62] B. Mirchevska, M. Blum, L. Louis, J. Boedecker, and M. Werling, "Reinforcement learning for autonomous maneuvering in highway scenarios," in *Workshop Driv. Assist. Syst. Auton. Driv.*, 2017, pp. 32–41.

[63] B. Mirchevska, C. Pek, M. Werling, M. Althoff, and J. Boedecker, "High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning," in *2018 21st Int. Conf. Intell. Transp. Syst. (ITSC)*. IEEE, 2018, pp. 2156–2162.

[64] X. Wang, "Ensuring safety of learning-based motion planners using control barrier functions," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 4773–4780, 2022.

[65] J. Garcia and F. Fernández, "Safe exploration of state and action spaces in reinforcement learning," *J. Artif. Intell. Res.*, vol. 45, pp. 515–564, 2012.

[66] J. García and D. Shafie, "Teaching a humanoid robot to walk faster through safe reinforcement learning," *Eng. Appl. Artif. Intell.*, vol. 88, p. 103360, 2020.

[67] T. J. Perkins and A. G. Barto, "Lyapunov design for safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 3, no. Dec, pp. 803–832, 2002.

[68] T.-H. Pham, G. De Magistris, and R. Tachibana, "Optlayer-practical constrained optimization for deep reinforcement learning in the real world," in *2018 IEEE Int. Conf. Robot. Autom (ICRA)*. IEEE, 2018, pp. 6236–6243.

[69] D. Slack, Y. Chow, B. Dai, and N. Wichers, "Safer: Data-efficient and safe reinforcement learning via skill acquisition," *arXiv:2202.04849*, 2022.

[70] G. Thomas, Y. Luo, and T. Ma, "Safe reinforcement learning by imagining the near future," *Adv. Neur. In.*, vol. 34, 2021.

[71] A. Mandhane, A. Zhernov, M. Rauh, C. Gu, M. Wang, F. Xue, W. Shang, D. Pang, R. Claus, C.-H. Chiang *et al.*, "Muzero with self-competition for rate control in vp9 video compression," *arXiv:2202.06626*, 2022.

[72] H. Li, Z. Wan, and H. He, "Constrained ev charging scheduling based on safe deep reinforcement learning," *IEEE Trans. Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2019.

[73] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, "Ai safety gridworlds," *arXiv:1711.09883*, 2017.

[74] A. Ray, J. Achiam, and D. Amodei, "Benchmarking safe exploration in deep reinforcement learning," *arXiv:1910.01708*, vol. 7, p. 1, 2019.

[75] S. Gu, J. G. Kuba, Y. Chen, Y. Du, L. Yang, A. Knoll, and Y. Yang, "Safe multi-agent reinforcement learning for multi-robot control," *Artif. Intell.*, vol. 319, p. 103905, 2023.

[76] J. J. Thomson, "Killing, letting die, and the trolley problem," *The monist*, vol. 59, no. 2, pp. 204–217, 1976.

[77] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, Y. Yang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," *arXiv:2205.10330*, 2022.

[78] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999.

[79] F. J. Beutler and K. W. Ross, "Optimal policies for controlled markov chains with a constraint," *J. Math. Anal. Appl.*, vol. 112, no. 1, pp. 236–252, 1985.

[80] L. Kallenberg, *Linear programming and finite Markovian control problems*, 01 1983, vol. 148.

[81] K. W. Ross, "Randomized and past-dependent policies for markov decision processes with multiple constraints," *Oper. Res.*, vol. 37, no. 3, pp. 474–477, 1989.

[82] K. W. Ross and R. Varadarajan, "Markov decision processes with sample path constraints: the communicating case," *Oper. Res.*, vol. 37, no. 5, pp. 780–790, 1989.

[83] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[84] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annu. Rev. Control Robot. Auton. Syst.*, vol. 5, 2021.

[85] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 1437–1480, 2015.

[86] Y. Kim, R. Allmendinger, and M. López-Ibáñez, "Safe learning and optimization techniques: Towards a survey of the state of the art," in *Int. Work. Found. Trustw. AI Integr. Learn. Optim. Reason.* Springer, 2020, pp. 123–139.

[87] Y. Liu, A. Halev, and X. Liu, "Policy learning with constraints in model-free reinforcement learning: A survey," in *Proc. the Thirtieth Int. Jt. Conf. Artif. Intell.*, 2021.

[88] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *Conf. Decis. Control. (CDC)*. IEEE, 2018, pp. 6059–6066.

[89] A. Gahlawat, P. Zhao, A. Patterson, N. Hovakimyan, and E. Theodorou, "L1-gp: L1 adaptive control with bayesian learning," in *Learn. Dyn. Control.* PMLR, 2020, pp. 826–837.

[90] A. von Rohr, M. Neumann-Brosig, and S. Trimpe, "Probabilistic robust linear quadratic regulators with gaussian processes," in *Learn. Dyn. Control.* PMLR, 2021, pp. 324–335.

[91] K. Regan and C. Boutilier, "Regret-based reward elicitation for markov decision processes," in *Proc. the Twenty-Fifth Conf. Uncertain. Artif. Intell.*, 2009, pp. 444–451.

[92] A. Basu, T. Bhattacharyya, and V. S. Borkar, "A learning algorithm for risk-sensitive cost," *Math. Oper. Res.*, vol. 33, no. 4, pp. 880–898, 2008.

[93] V. S. Borkar, "Q-learning for risk-sensitive control," *Math. Oper. Res.*, vol. 27, no. 2, pp. 294–311, 2002.

[94] M. Heger, "Consideration of risk in reinforcement learning," in *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 105–111.

[95] R. A. Howard and J. E. Matheson, "Risk-sensitive markov decision processes," *Management science*, vol. 18, no. 7, pp. 356–369, 1972.

[96] Y. Kadota, M. Kurano, and M. Yasuda, "Discounted markov decision processes with utility constraints," *Comput. Math. Appl.*, vol. 51, no. 2, pp. 279–284, 2006.

[97] B. Lütjens, M. Everett, and J. P. How, "Safe reinforcement learning with model uncertainty estimates," in *2019 Int. Conf. Robot. Autom. (ICRA)*. IEEE, 2019, pp. 8662–8668.

[98] A. Nilim and L. El Ghaoui, "Robust control of markov decision processes with uncertain transition matrices," *Oper. Res.*, vol. 53, no. 5, pp. 780–798, 2005.

[99] M. Sato, H. Kimura, and S. Kobayashi, "Td algorithm for the variance of return and mean-variance reinforcement learning," *Trans. Jpn. Soc. Artif. Intell.*, vol. 16, no. 3, pp. 353–362, 2001.

[100] A. Tamar, D. Di Castro, and S. Mannor, "Policy gradients with variance related risk criteria," in *Proc. the 29th International Coference on Int. Conf. Mach. Learn.*, 2012, pp. 1651–1658.

[101] P. Abbeel, A. Coates, and A. Y. Ng, "Autonomous helicopter aerobatics through apprenticeship learning," *Int. J. Robot. Res.*, vol. 29, no. 13, pp. 1608–1639, 2010.

[102] J. A. Clouse and P. E. Utgoff, "A teaching method for reinforcement learning," in *Machine learning proceedings 1992*. Elsevier, 1992, pp. 92–101.

[103] A. Geramifard, J. Redding, and J. P. How, "Intelligent cooperative control architecture: a framework for performance improvement using safe learning," *J. Intell. Robot. Syst.*, vol. 72, no. 1, pp. 83–103, 2013.

[104] J. Tang, A. Singh, N. Goehausen, and P. Abbeel, "Parameterized maneuver learning for autonomous helicopter flight," in *2010 IEEE Int. Conf. Robot. Autom.* IEEE, 2010, pp. 1142–1148.

[105] A. L. Thomaz and C. Breazeal, "Reinforcement learning with human teachers: evidence of feedback and guidance with implications for learning performance," in *Proc. the 21st Natl. Conf. Artif. Intell., Vol 1*, 2006, pp. 1000–1005.

[106] V. Khattar, Y. Ding, B. Sel, J. Lavaei, and M. Jin, "A cmdp-within-online framework for meta-safe reinforcement learning," in *ICLR*, 2022.

[107] Q. Bai, A. S. Bedi, M. Agarwal, A. Koppel, and V. Aggarwal, "Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 4, 2022, pp. 3682–3689.

[108] Y. Efroni, S. Mannor, and M. Pirotta, "Exploration-exploitation in constrained mdps," *arXiv:2003.02189*, 2020.

[109] T. Liu, R. Zhou, D. Kalathil, P. R. Kumar, and C. Tian, "Learning policies with zero or bounded constraint violation for constrained mdps," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2021.

[110] H. Wei, X. Liu, and L. Ying, "A provably-efficient model-free algorithm for constrained markov decision processes," *arXiv:2106.01577*, 2021.

[111] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," *arXiv:1805.11074*, 2018.

[112] V. S. Borkar, "An actor-critic algorithm for constrained markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[113] A. Mas-Colell, M. D. Whinston, J. R. Green *et al.*, *Microeconomic theory*. Oxford university press New York, 1995, vol. 1.

[114] P. Milgrom and I. Segal, "Envelope theorems for arbitrary choice sets," *Econometrica*, vol. 70, no. 2, pp. 583–601, 2002.

[115] M. Yu, Z. Yang, M. Kolar, and Z. Wang, "Convergent policy optimization for safe reinforcement learning," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2019.

[116] A. Liu, V. K. Lau, and B. Kananian, "Stochastic successive convex approximation for non-convex constrained stochastic optimization," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4189–4203, 2019.

[117] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 6070–6120, 2017.

[118] O. Bardou, N. Frikha, and G. Pages, "Computing var and cvar using stochastic approximation and adaptive unconstrained importance sampling," 2009.

[119] A. Tamar, Y. Glassner, and S. Mannor, "Policy gradients beyond expectations: Conditional value-at-risk," *arXiv:1404.3862*, 2014.

[120] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, vol. 33, 2020.

[121] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[122] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Adv. Neur. In.*, vol. 30, 2017.

[123] M. Zanon and S. Gros, "Safe reinforcement learning using robust mpc," *IEEE Trans. Autom. Control*, vol. 66, no. 8, pp. 3638–3652, 2020.

[124] A. Aswani, H. Gonzalez, S. S. Sastry, and C. Tomlin, "Provably safe and robust learning-based model predictive control," *Automatica J. IFAC*, vol. 49, no. 5, pp. 1216–1226, 2013.

[125] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica J. IFAC*, vol. 36, no. 6, pp. 789–814, 2000.

[126] J. F. Fisac, A. K. Akametalu, M. N. Zeilinger, S. Kaynama, J. Gillula, and C. J. Tomlin, "A general safety framework for learning-based control in uncertain robotic systems," *IEEE Trans. Autom. Control*, vol. 64, no. 7, pp. 2737–2752, 2018.

[127] H. Ma, J. Chen, S. Eben, Z. Lin, Y. Guan, Y. Ren, and S. Zheng, "Model-based constrained reinforcement learning using generalized control barrier function," in *IROS 2021*. IEEE, 2021, pp. 4552–4559.

[128] Y. Emam, G. Notomista, P. Glotfelter, Z. Kira, and M. Egerstedt, "Safe reinforcement learning using robust control barrier functions," *IEEE Robot. Autom. Lett.*, 2022.

[129] M. H. Cohen and C. Belta, "Safe exploration in model-based reinforcement learning using control barrier functions," *Automatica*, vol. 147, p. 110684, 2023.

[130] Z. Marvi and B. Kiumarsi, "Safe reinforcement learning: A control barrier function optimization approach," *Int. J. Robust Nonlinear Control*, vol. 31, no. 6, pp. 1923–1940, 2021.

[131] Y. Emam, P. Glotfelter, and M. Egerstedt, "Robust barrier functions for a fully autonomous, remotely accessible swarm-robotics testbed," in *CDC 2019*. IEEE, 2019, pp. 3984–3990.

[132] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. Conf. Mach. Learn.* PMLR, 2018, pp. 1861–1870.

[133] D. Panagou, D. M. Stipanović, and P. G. Voulgaris, "Distributed coordination control for multi-robot networks using lyapunov-like barrier functions," *IEEE TAC*, vol. 61, no. 3, pp. 617–632, 2015.

[134] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2, no. 3.

[135] F. Berkenkamp and A. P. Schoellig, "Safe and robust learning control with gaussian processes," in *2015 European Control Conference (ECC)*. IEEE, 2015, pp. 2496–2501.

[136] M. Deisenroth and C. E. Rasmussen, "Pilco: A model-based and data-efficient approach to policy search," in *Proc. the 28th Int. Conf. Mach. Learn. (ICML-11)*. Citeseer, 2011, pp. 465–472.

[137] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1889–1897.

[138] J. Peters, K. Mülling, and Y. Altun, "Relative entropy policy search." in *AAAI*, 2010, pp. 1607–1612.

[139] N. Wagener, B. Boots, and C.-A. Cheng, "Safe reinforcement learning using advantage-based intervention," in *Int. Conf. Mach. Learn.*, 2021.

[140] S. Sohn, S. Lee, J. Choi, H. van Seijen, M. Fatemi, and H. Lee, "Shortest-path constrained reinforcement learning for sparse reward tasks," in *Int. Conf. Mach. Learn.*, 2021.

[141] A. Wachi and Y. Sui, "Safe reinforcement learning in constrained markov decision processes," in *Int. Conf. Mach. Learn.* PMLR, 2020, pp. 9797–9806.

[142] Y. Liu, J. Ding, and X. Liu, "Ipo: Interior-point policy optimization under constraints," in *Proc. the AAAI Conf. Artif. Intell.*, vol. 34, no. 04, 2020, pp. 4940–4947.

[143] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[144] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, "Risk-sensitive and robust decision-making: a cvar optimization approach," *Adv. Neur. In.*, vol. 28, 2015.

[145] K. Brantley, M. Dudik, T. Lykouris, S. Miryoosefi, M. Simchowitz, A. Slivkins, and W. Sun, "Constrained episodic reinforcement learning in concave-convex and knapsack settings," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2020.

[146] D. Blackwell, "An analog of the minimax theorem for vector payoffs." *Pacific J. Math.*, vol. 6, no. 1, pp. 1–8, 1956.

[147] S. Miryoosefi, K. Brantley, H. Daumé III, M. Dudík, and R. Schapire, "Reinforcement learning with convex constraints," in *Adv. Neural Inf. Process. Syst.(NeurIPS)*, 2019.

[148] S. Paternain, M. Calvo-Fullana, L. F. Chamon, and A. Ribeiro, "Safe policies for reinforcement learning via primal-dual methods," *arXiv:1911.09101*, 2019.

[149] I. Panageas, G. Piliouras, and X. Wang, "First-order methods almost always avoid saddle points: The case of vanishing step-sizes," *Adv. Neur. In.*, vol. 32, 2019.

[150] A. Ghosh, X. Zhou, and N. Shroff, "Provably efficient model-free constrained rl with linear function approximation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 13 303–13 315, 2022.

[151] C. Jin, Z. Yang, Z. Wang, and M. I. Jordan, "Provably efficient reinforcement learning with linear function approximation," in *Proc. 33rd COLT*, ser. Proceedings of Machine Learning Research, J. Abernethy and S. Agarwal, Eds., vol. 125. PMLR, 09–12 Jul 2020, pp. 2137–2143. [Online]. Available: https://proceedings.mlr.press/v125/jin20a.html

[152] N. Xiong, Y. Du, and L. Huang, "Provably safe reinforcement learning with step-wise violation constraints," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[153] M. G. Azar, I. Osband, and R. Munos, "Minimax regret bounds for reinforcement learning," in *Int. Conf. Mach. Learn.* PMLR, 2017, pp. 263–272.

[154] Y. Yang, Y. Jiang, Y. Liu, J. Chen, and S. E. Li, "Model-free safe reinforcement learning through neural barrier certificate," *IEEE RAL*, vol. 8, no. 3, pp. 1295–1302, 2023.

[155] T. L. Vu, S. Mukherjee, R. Huang, and Q. Huang, "Barrier function-based safe reinforcement learning for emergency control of power systems," in *CDC 2021*. IEEE, 2021, pp. 3652–3657.

[156] Z. Marvi and B. Kiumarsi, "Safe off-policy reinforcement learning using barrier functions," in *ACC 2020*. IEEE, 2020, pp. 2176–2181.

[157] B. Zhang, Y. Zhang, L. Frison, T. Brox, and J. Bödecker, "Constrained reinforcement learning with smoothed log barrier function," *arXiv:2403.14508*, 2024.

[158] R. Cheng, G. Orosz, R. M. Murray, and J. W. Burdick, "End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, 2019, pp. 3387–3395.

[159] S. Liu, C. Liu, and J. Dolan, "Safe control under input limits with neural control barrier functions," in *CoRL*. PMLR, 2023, pp. 1970–1980.

[160] F. B. Mathiesen, S. C. Calvert, and L. Laurenti, "Safety certification for stochastic systems via neural barrier functions," *IEEE Control Syst. Lett.*, vol. 7, pp. 973–978, 2022.

[161] Z. Qin, Y. Chen, and C. Fan, "Density constrained reinforcement learning," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8682–8692.

[162] B. Dai, A. Shaw, N. He, L. Li, and L. Song, "Boosting the actor with dual critic," in *Int. Conf. Learn. Represent.*, 2018.

[163] O. Nachum, Y. Chow, B. Dai, and L. Li, "Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections," *Adv. Neur. In.*, vol. 32, 2019.

[164] O. Nachum and B. Dai, "Reinforcement learning via fenchel-rockafellar duality," *arXiv:2001.01866*, 2020.

[165] Z. Tang, Y. Feng, L. Li, D. Zhou, and Q. Liu, "Doubly robust bias reduction in infinite horizon off-policy estimation," in *Int. Conf. Learn. Represent.*, 2019.

[166] G. Dalal, K. Dvijotham, M. Vecerik, T. Hester, C. Paduraru, and Y. Tassa, "Safe exploration in continuous action spaces," *arXiv:1801.08757*, 2018.

[167] Y. Chen, A. W. Singletary, and A. D. Ames, "Density functions for guaranteed safety on robotic systems," in *2020 Am. Control Conf. (ACC)*. IEEE, 2020, pp. 3199–3204.

[168] M. Calvo-Fullana, S. Paternain, L. F. Chamon, and A. Ribeiro, "State augmented constrained reinforcement learning: Overcoming the limitations of learning with rewards," in *Int. Conf. Mach. Learn.*, 2021.

[169] M. Hasanbeig, A. Abate, and D. Kroening, "Cautious reinforcement learning with logical constraints," in *Proc. the 19th Int. Conf. AAMAS*, 2020, pp. 483–491.

[170] A. Murugesan, M. Moghadamfalahi, and A. Chattopadhyay, "Formal methods assisted training of safe reinforcement learning agents," in *NFM 2019*. Springer, 2019, pp. 333–340.

[171] E. M. Clarke, T. A. Henzinger, H. Veith, R. Bloem *et al.*, *Handbook of model checking*. Springer, 2018, vol. 10.

[172] "Towards verifiable and safe model-free reinforcement learning." CEUR Workshop Proceedings, 2020.

[173] S. Sickert, J. Esparza, S. Jaax, and J. Křetínskỳ, "Limit-deterministic büchi automata for linear temporal logic," in *CAV*. Springer, 2016, pp. 312–332.

[174] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *In Proc. the Seventeenth Int. Conf. Mach. Learn.* Citeseer, 2000.

[175] C. Szepesvári, "Constrained MDPs and the reward hypothesis," http://readingsml.blogspot.com/2020/03/constrained-mdps-and-reward-hypothesis.html, 2020, Access on January 21, 2023.

[176] E. A. Feinberg, "Constrained discounted markov decision processes and hamiltonian cycles," *Mathematics of Operations Research*, vol. 25, no. 1, pp. 130–140, 2000.

[177] D. Ding, C.-Y. Wei, K. Zhang, and A. Ribeiro, "Last-iterate convergent policy gradient primal-dual methods for constrained mdps," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[178] F. J. V. Abad and V. Krishnamurthy, "Self learning control of constrained markov decision processes–a gradient approach," *Les Cahiers du GERAD ISSN*, vol. 711, p. 2440, 2003.

[179] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.

[180] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.

[181] D. Pollard, "Asymptopia: an exposition of statistical asymptotic theory. 2000," *URL http://www. stat. yale. edu/pollard/Books/Asymptopia*, 2000.

[182] A. M. Thompson and W. R. Cluett, "Stochastic iterative dynamic programming: A monte carlo approach to dual control," *Automatica J. IFAC*, vol. 41, no. 5, pp. 767–778, 2005.

[183] L. Yang, M. Shi, Q. Zheng, W. Meng, and G. Pan, "A unified approach for multi-step temporal-difference learning with eligibility traces in reinforcement learning," in *Proc. the Twenty-Seventh Int. Jt. Conf. Artif. Intell., IJCAI-18*, 2018, pp. 2984–2990.

[184] S. Bhatnagar and K. Lakshmanan, "An online actor–critic algorithm with function approximation for constrained markov decision processes," *J. Optim. Theory Appl.*, vol. 153, no. 3, pp. 688–708, 2012.

[185] S. Vaswani, L. Yang, and C. Szepesvári, "Near-optimal sample complexity bounds for constrained mdps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3110–3122, 2022.

[186] F. V. Abad, V. Krishnamurthy, K. Martin, and I. Baltcheva, "Self learning control of constrained markov chains-a gradient approach," in *Proceedings of the 41st IEEE Conference on Decision and Control, 2002.*, vol. 2. IEEE, 2002, pp. 1940–1945.

[187] Y. Chow, M. Ghavamzadeh, L. Janson, and M. Pavone, "Risk-constrained reinforcement learning with percentile risk criteria," *Journal of Machine Learning Research*, vol. 18, no. 167, pp. 1–51, 2018.

[188] T. Li, Z. Guan, S. Zou, T. Xu, Y. Liang, and G. Lan, "Faster algorithm and sharper analysis for constrained markov decision process," *Operations Research Letters*, vol. 54, p. 107107, 2024.

[189] A. HasanzadeZonuzy, D. Kalathil, and S. Shakkottai, "Model-based reinforcement learning for infinite-horizon discounted constrained markov decision processes," *IJCAI 2021*, 2021.

[190] D. Bertsekas, *Convex optimization algorithms*. Athena Scientific, 2015.

[191] M. El Chamie, Y. Yu, and B. Açıkmeşe, "Convex synthesis of randomized policies for controlled markov chains with density safety upper bound constraints," in *2016 Am. Control Conf. (ACC)*. IEEE, 2016, pp. 6290–6295.

[192] Z. Gábor, Z. Kalmár, and C. Szepesvári, "Multi-criteria reinforcement learning." in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 98, 1998, pp. 197–205.

[193] L. Yang, Q. Zheng, and G. Pan, "Sample complexity of policy gradient finding second-order stationary points," in *AAAI*, 2021.

[194] E. Süli and D. F. Mayers, *An introduction to numerical analysis*. Cambridge university press, 2003.

[195] M. Slater, "Lagrange multipliers revisited: a contribution to non-linear programming," 1950.

[196] T. Lattimore and M. Hutter, "Pac bounds for discounted mdps," in *Proc. the 23rd International Conference on Algorithmic Learning Theory*, ser. ALT'12, 2012, p. 320–334.

[197] M. G. Azar, R. Munos, and H. J. Kappen, "Minimax pac bounds on the sample complexity of reinforcement learning with a generative model," *Machine learning*, vol. 91, no. 3, pp. 325–349, 2013.

[198] E. Altman, "Constrained markov decision processes," Ph.D. dissertation, INRIA, 1995.

[199] L. Yang, J. Ji, J. Dai, L. Zhang, B. Zhou, P. Li, Y. Yang, and G. Pan, "Constrained update projection approach to safe policy optimization," *Adv. Neur. In.*, vol. 35, pp. 9111–9124, 2022.

[200] Z. Zhou, H. Wei, and L. Ying, "Model-free, regret-optimal best policy identification in online cmdps," *arXiv preprint arXiv:2309.15395*, 2023.

[201] G. Tesauro, "Td-gammon, a self-teaching backgammon program, achieves master-level play," *Neural computation*, vol. 6, no. 2, pp. 215–219, 1994.

[202] G. Tesauro *et al.*, "Temporal difference learning and td-gammon," *Commun. ACM*, vol. 38, no. 3, pp. 58–68, 1995.

[203] R. H. Crites and A. G. Barto, "Elevator group control using multiple reinforcement learning agents," *Machine learning*, vol. 33, no. 2, pp. 235–262, 1998.

[204] J. N. Tsitsiklis and B. Van Roy, "Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives," *IEEE Trans. Autom. Control*, vol. 44, no. 10, pp. 1840–1851, 1999.

[205] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Adv. Neur. In.*, vol. 1, 1988.

[206] S. Gu, C. Zhou, Y. Wen, C. Xiao, Z. Du, and L. Huang, "Path search of unmanned surface vehicle based on topological location," *Navigation of China*, vol. 42, no. 02, pp. 52–58, 2019.

[207] C. Zhou, S. Gu, Y. Wen, Z. Du, C. Xiao, L. Huang, and M. Zhu, "Motion planning for an unmanned surface vehicle based on topological position maps," *Ocean Eng.*, vol. 198, p. 106798, 2020.

[208] S. Gu, C. Zhou, Y. Wen, C. Xiao, and A. Knoll, "Motion planning for an unmanned surface vehicle with wind and current effects," *J. mar. sci. eng.*, vol. 10, no. 3, p. 420, 2022.

[209] S. Gu, C. Zhou, Y. Wen, X. Zhong, M. Zhu, C. Xiao, and Z. Du, "A motion planning method for unmanned surface vehicle in restricted waters," *Proc. Inst. Mech. Eng*, vol. 234, no. 2, pp. 332–345, 2020.

[210] C. Zhou, S. Gu, Y. Wen, Z. Du, C. Xiao, L. Huang, and M. Zhu, "The review unmanned surface vehicle path planning: Based on multi-modality constraint," *Ocean Eng.*, vol. 200, p. 107043, 2020.

[211] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, "Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization," in *2020 IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*. IEEE, 2020, pp. 1–7.

[212] A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, and P. Tabuada, "Control barrier functions: Theory and applications," in *2019 18th European control conference (ECC)*. IEEE, 2019, pp. 3420–3431.

[213] Z. Cao, S. Xu, X. Jiao, H. Peng, and D. Yang, "Trustworthy safety improvement for autonomous driving using reinforcement learning," *Transp. Res. C: Emerg*, vol. 138, p. 103656, 2022.

[214] P. G. Gipps, "A behavioural car-following model for computer simulation," *Transport Res. B-Meth.*, vol. 15, no. 2, pp. 105–111, 1981.

[215] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, "A survey of monte carlo tree search methods," *IEEE Trans Comput Intell AI Games*, vol. 4, no. 1, pp. 1–43, 2012.

[216] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3, pp. 279–292, 1992.

[217] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, no. 1, pp. 3–42, 2006.

[218] C. Pek, P. Zahn, and M. Althoff, "Verifying the safety of lane change maneuvers of self-driving vehicles based on formalized traffic rules," in *Proc. of the IIEEE Intell. Veh. Symp.*, 2017.

[219] D. Kamran, T. Engelgeh, M. Busch, J. Fischer, and C. Stiller, "Minimizing safety interference for safe and comfortable automated driving with distributional reinforcement learning," in *2021 IEEE Int. Conf. Intell. Robots Syst. (IROS)*. IEEE, 2021, pp. 1236–1243.

[220] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Int. Conf. Mach. Learn.* PMLR, 2014, pp. 1278–1286.

[221] 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," *Int. Conf. Learn. Represent. (ICLR)*, 2016.

[222] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st Int. Conf. Intell. Transp. Syst. (ITSC)*. IEEE, 2018, pp. 2575–2582.

[223] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st Int. Conf. Intell. Transp. Syst. (ITSC)*. IEEE, 2018, pp. 2118–2125.

[224] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions," *arXiv:2112.11561*, 2021.

[225] J. Kober and J. Peters, "Policy search for motor primitives in robotics," in *Twenty-Second Annu. Conf. Neural Inf. Process. Syst. (NIPS 2008)*. Curran, 2009, pp. 849–856.

[226] L. Yang, Z. Huang, F. Lei, Y. Zhong, Y. Yang, C. Fang, S. Wen, B. Zhou, and Z. Lin, "Policy representation via diffusion probability model for reinforcement learning," *arXiv:2305.13122*, 2023.

[227] A. Singh, H. Liu, G. Zhou, A. Yu, N. Rhinehart, and S. Levine, "Parrot: Data-driven behavioral priors for reinforcement learning," in *Int. Conf. Learn. Represent.*, 2020.

[228] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[229] R. Kalman and J. Bertram, "Control system analysis and design via the second method of lyapunov:(i) continuous-time systems (ii) discrete time systems," *IRE Trans. Autom. Control*, vol. 4, no. 3, pp. 112–112, 1959.

[230] R. Sepulchre, M. Jankovic, and P. V. Kokotovic, *Constructive nonlinear control*. Springer Science & Business Media, 2012.

[231] P. Liu, D. Tateo, H. B. Ammar, and J. Peters, "Robot reinforcement learning on the constraint manifold," in *Conference on Robot Learning*. PMLR, 2022, pp. 1357–1366.

[232] A. Singh, Y. Halpern, N. Thain, K. Christakopoulou, E. Chi, J. Chen, and A. Beutel, "Building healthy recommendation sequences for everyone: A safe reinforcement learning approach," in *FAccTRec Workshop*, 2020.

[233] L. Xiao, Y. Ding, J. Huang, S. Liu, Y. Tang, and H. Dai, "Uav anti-jamming video transmissions with qoe guarantee: A reinforcement learning-based approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5933–5947, 2021.

[234] X. Lu, L. Xiao, G. Niu, X. Ji, and Q. Wang, "Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 732–743, 2022.

[235] K. Dunlap, M. Mote, K. Delsing, and K. L. Hobbs, "Run time assured reinforcement learning for safe satellite docking," in *AIAA SCITECH 2022 Forum*, 2022, p. 1853.

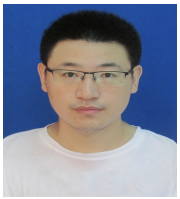
[236] L. Xiao, X. Lu, T. Xu, X. Wan, W. Ji, and Y. Zhang, "Reinforcement learning-based mobile offloading for edge computing against jamming and interference," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6114–6126, 2020.

[237] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the cvar via sampling," in *Twenty-Ninth AAAI Conf. Artif. Intell.*, 2015.

[238] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel *et al.*,

"Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.
[239] Y. Wang, S. Inguva, and B. Adsumilli, "Youtube ugc dataset for video compression research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.
[240] F. L. Da Silva, G. Warnell, A. H. R. Costa, and P. Stone, "Agents teaching agents: a survey on inter-agent transfer learning," *AAMAS*, vol. 34, no. 1, pp. 1–17, 2020.
[241] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recogn.*, vol. 77, pp. 354–377, 2018.
[242] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv:1606.01540*, 2016.
[243] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE Int. Conf. Intell. Robots Syst.* IEEE, 2012, pp. 5026–5033.
[244] Z. Yuan, A. W. Hall, S. Zhou, L. Brunke, M. Greeff, J. Panerati, and A. P. Schoellig, "safe-control-gym: a unified benchmark suite for safe learning-based control and reinforcement learning," *arXiv:2109.06325*, 2021.
[245] J. Buchli, F. Farshidian, A. Winkler, T. Sandy, and M. Giftthaler, "Optimal and learning control for autonomous robots," *arXiv:1708.09342*, 2017.
[246] L. Hewing, J. Kabzan, and M. N. Zeilinger, "Cautious model predictive control using gaussian process regression," *IEEE Trans. Control Syst. Technol.*, vol. 28, no. 6, pp. 2736–2743, 2019.
[247] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. Conf. Mach. Learn. (ICML)*, 2018.
[248] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv:1707.06347*, 2017.
[249] B. Peng, T. Rashid, C. A. S. de Witt, P.-A. Kamienny, P. H. Torr, W. Böhmer, and S. Whiteson, "Facmac: Factored multi-agent centralised policy gradients," *arXiv:2003.06709*, 2020.
[250] M. A. Zanger, K. Daaboul, and J. M. Zöllner, "Safe continuous control with constrained model-based policy optimization," 2021.
[251] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín, "robosuite: A modular simulation framework and benchmark for robot learning," *arXiv:2009.12293*, 2020.
[252] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu based physics simulation for robot learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.
[253] A. Fickinger, S. Zhuang, D. Hadfield-Menell, and S. Russell, "Multi-principal assistance games," *arXiv:2007.09540*, 2020.
[254] E. Hughes, J. Z. Leibo, M. Phillips, K. Tuyls, E. Dueñez-Guzman, A. García Castañeda, I. Dunning, T. Zhu, K. McKee, R. Koster *et al.*, "Inequity aversion improves cooperation in intertemporal social dilemmas," *Adv. Neur. In.*, vol. 31, 2018.
[255] J. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multi-agent reinforcement learning in sequential social dilemmas," in *AAMAS*, vol. 16. ACM, 2017, pp. 464–473.
[256] P. Foot, "The problem of abortion and the doctrine of the double effect," *Oxford review*, vol. 5, 1967.
[257] T. Liu, R. Zhou, D. Kalathil, P. Kumar, and C. Tian, "Policy optimization for constrained mdps with provable fast global convergence," *arXiv preprint arXiv:2111.00552*, 2021.
[258] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

Shangding Gu is currently working as a postdoc at UC Berkeley, and he is a member of the Informatics 6-Chair of Robotics, Artificial Intelligence and Real-time Systems, Technical University of Munich. He is one of the organizers of the 1st International Safe Reinforcement Learning Workshop at IEEE MFI 2022. His main research interests include safe reinforcement learning and motion planning.



Long Yang is currently working as a postdoc in the Key Lab, working with Prof. Zhouchen Lin. Before working in Peking University, He obtained his Ph.D. degree in School of Computer Science and Technology from Zhejiang University in 2021.

Yali Du is a Lecturer in AI at King's College London, and a Turing Fellow at The Alan Turing Institute. Her research aims to enable machines to exhibit cooperative and responsible behaviour in intelligent decision making tasks. She serves as the editors for Journal of AAMAS and IEEE Transactions on AI, Area Chair for NeurIPS 2024. She also serves in organising committee for AAMAS 2023 and NeurIPS 2024.

Guang Chen is a professor at Tongji University and a senior research associate (guest) at Technical University of Munich. His research interests include 3D vision, intelligent robotics and autonomous driving. He was awarded the program of Shanghai Rising Star 2021, and Shanghai S&T 35U35 2021, the National Distinguished Young Talents 2023. He serves as an Associate Editor for several international journals. He is the program chair of IEEE MFI 2022.

Florian Walter received his Master's degree in informatics with high distinction from Technische Universität München. During his Master studies, he completed an internship in the automotive industry and was a visiting student researcher in the Artificial Intelligence Laboratory at Stanford University where he worked in the field of online trajectory generation for robotics.

Jun Wang is a chair professor at University College London (UCL), and Founding Director of MSc Web Science and Big Data Analytics. He is also a co-founder and chief scientist at MediaGamma Ltd, a UCL start-up company focusing on AI for intelligent audience decision-making. His main research interests are in AI and intelligent systems, including reinforcement learning and deep generative models.

Alois Knoll (IEEE Fellow) is a professor at the Department of Informatics, TU Munich. From 2004 to 2006, he was Executive Director of the Institute of Computer Science at TUM. He was also on the board of directors of the Central Institute of Medical Technology at TUM. His research interests include cognitive, and sensor-based robotics, multi-agent systems, data fusion, adaptive systems, multimedia information retrieval, and model-driven development of embedded systems.