electricity

MDPI

Review

Deep Reinforcement Learning for Resilient Power and Energy Systems: Progress, Prospects, and Future Avenues

Mukesh Gautam

Power and Energy Systems Department, Idaho National Laboratory, Idaho Falls, ID 83415, USA; mukesh.gautam@nevada.unr.edu

Abstract: In recent years, deep reinforcement learning (DRL) has garnered substantial attention in the context of enhancing resilience in power and energy systems. Resilience, characterized by the ability to withstand, absorb, and quickly recover from natural disasters and human-induced disruptions, has become paramount in ensuring the stability and dependability of critical infrastructure. This comprehensive review delves into the latest advancements and applications of DRL in enhancing the resilience of power and energy systems, highlighting significant contributions and key insights. The exploration commences with a concise elucidation of the fundamental principles of DRL, highlighting the intricate interplay among reinforcement learning (RL), deep learning, and the emergence of DRL. Furthermore, it categorizes and describes various DRL algorithms, laying a robust foundation for comprehending the applicability of DRL. The linkage between DRL and power system resilience is forged through a systematic classification of DRL applications into five pivotal dimensions: dynamic response, recovery and restoration, energy management and control, communications and cybersecurity, and resilience planning and metrics development. This structured categorization facilitates a methodical exploration of how DRL methodologies can effectively tackle critical challenges within the domain of power and energy system resilience. The review meticulously examines the inherent challenges and limitations entailed in integrating DRL into power and energy system resilience, shedding light on practical challenges and potential pitfalls. Additionally, it offers insights into promising avenues for future research, with the aim of inspiring innovative solutions and further progress in this vital domain.

Keywords: communications and cybersecurity; deep learning; deep reinforcement learning; dynamic response; energy management and control; power and energy system resilience; resilience review

Citation: Gautam, M. Deep Reinforcement Learning for Resilient Power and Energy Systems: Progress, Prospects, and Future Avenues. *Electricity* **2023**, *4*, 336–380. https://doi.org/10.3390/electricity4040020

Academic Editor: Andreas Sumper

Received: 7 October 2023
Revised: 4 November 2023
Accepted: 16 November 2023
Published: 1 December 2023

1. Introduction

In recent decades, the frequency of extreme events, including natural disasters such as droughts, heatwaves, floods, hurricanes, wildfires, earthquakes, and winter storms, as well as man-made disruptions like cyber-attacks, has notably increased [1]. Research conducted by the U.S. Energy Information Administration has confirmed a significant rise in the occurrence of weather-related extreme events in the United States between 1992 and 2012 [2,3]. Figure 1 provides a visual representation of the notable increase in weather-related extreme events in the United States, characterized by economic damages exceeding one billion dollars. The data, sourced from the National Oceanic and Atmospheric Administration [4], spans the years from 1980 to 2022. Notably, the year 2022 witnessed the occurrence of 18 severe weather-related disasters within the United States, each resulting in economic losses exceeding the billion-dollar threshold. Extreme events are impacting regions across the globe, not limited to the United States. Examples of such events include a severe storm in Australia in 2016, a windstorm in Canada in 2015, and the 2016 tornado of Jiangsu Province in China [5]. These events have had a devastating impact on critical power and energy system components, resulting in extensive and prolonged power disruptions. The

significant increase in the frequency and economic impact of extreme weather events, as depicted in Figure 1, underscores the pressing need for resilient power and energy systems.

The growing influence of global warming, exemplified by the increasing prevalence of hurricanes and other natural disasters, has heightened the importance of power and energy system resilience. While power infrastructure has historically focused on reliability, aiming to withstand known threats and ensure uninterrupted power supply, the rise in extreme weather events presents a significant challenge [6]. Between 2003 and 2012, approximately 679 large-scale power outages in the United States were attributed to extreme weather, each affecting a minimum of 50,000 customers and resulting in an annual economic loss exceeding USD 18 billion [3]. These recurring and disruptive events underscore the limitations of current power facilities in effectively mitigating their impact [7]. While the likelihood of extreme natural events may be relatively low, the severity of their impact is indisputable. Consequently, there is an urgent need to enhance power and energy systems' resilience to withstand and recover from such events.

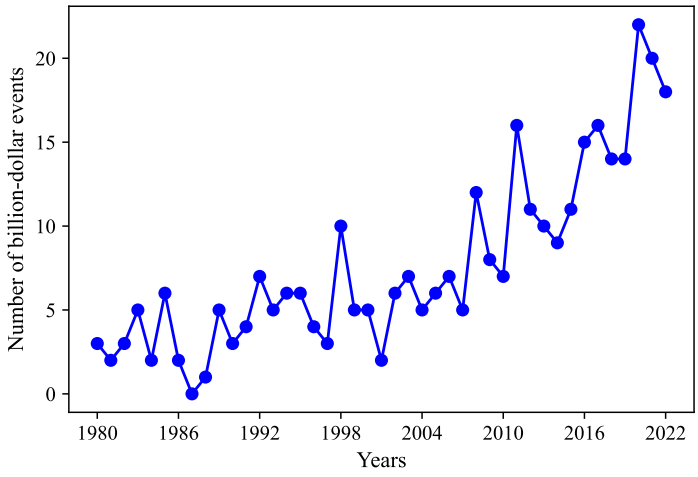

Figure 1. Occurrence of climate disasters in the United States from 1980 to 2022 [8].

To enhance the reliability and resilience of power and energy systems, various analytical and population-based heuristic approaches employing dynamic modeling and optimization techniques have been used in the literature. For instance, previous work [9] has utilized dynamic modeling and classical optimization to address threats like wildfires to power distribution networks, focusing on resilience with renewable energy resources and evaluated on a 33-node distribution system. Another study [10] analyzed microgrid operation using a multicarrier energy hub, considering various energy carriers and resources to reduce environmental impact and operational costs while enhancing resilience and flexibility. Similarly, methodologies based on graph theory and cooperative game theory [11,12] were introduced to determine optimal locations and sizes of movable energy resources for power distribution system resilience. Additionally, an optimal energy storage sizing method [13] aimed to improve reliability and resilience in networked microgrids through bi-level optimization, showcasing the benefits of microgrid interconnection in enhancing both reliability and resilience during grid outages compared to non-networked microgrids. Furthermore, addressing the impact of unfavorable weather and natural disasters, an article [14] emphasized evaluating network resilience and proposed a cost-effective method involving tie-lines for power restoration, outperforming other optimization techniques. In [15], a labor-economics-based framework, employing contract theory, was introduced to model interactions in modern smart grid systems, focusing on the microgrid operator and prosumers, with the potential to establish optimal personalized contracts for energy selling and purchasing, thereby contributing to resilient energy management and control in smart grids while satisfying the involved parties' profit and requirements. The evolutionary swarm algorithm (ESA) [16] was introduced for reliability assessment, offering advan-

tages in accuracy, computational efficiency, and precision over other heuristic approaches. However, these analytical and heuristic techniques have limitations, including modeling inaccuracies due to data scarcity and scalability challenges [17]. Additionally, these methods require repeated calculations when applied to new scenarios [18]. To address these shortcomings, recent advances have introduced deep reinforcement learning (DRL)-based approaches for enhancing the resilience of power and energy systems.

Within power and energy systems, DRL, a combination of reinforcement learning (RL) and deep learning, has emerged as an attractive alternative for conventional analytical and heuristic methods, offering solutions to their inherent shortcomings. Similar to other learning-driven methodologies, DRL makes use of past experiences to inform decision making. In [19], a real-time dynamic optimal energy management system for microgrids utilized DRL, specifically the proximal policy optimization (PPO) technique, to enhance efficiency and stability while integrating renewable energy sources. This approach showcased superior computational accuracy and efficiency compared to conventional mathematical programming or heuristic strategies. Additionally, in [20], a new energy management approach employed DRL within a Markov decision process (MDP) framework to minimize daily operating costs without the need for explicit uncertainty prediction, highlighting its effectiveness with real power-grid data. Furthermore, an innovative DRL-based Volt-VAR control and optimization method was introduced in [21], showcasing its effectiveness in improving voltage profiles, reducing power losses, and optimizing operational costs on various test cases. References [22,23] implemented Volt-VAR optimization in distribution grids with high DER penetration and volt-VAR control in active distribution systems, respectively, both leveraging DRL for efficient reactive power management. A DRL-based trusted collaborative computing has been proposed and analyzed in [24] for intelligent vehicle networks. A federated DRL-based approach for wind power forecasting has been proposed in [25], which is supposed to handle data sharing and privacy concerns. Beyond these aforementioned DRL applications, there is a growing landscape of wide-area applications in enhancing power system resilience, making this review paper a comprehensive exploration of DRL's increasing significance in this field.

Numerous review papers exist in the fields of power and energy systems, resilience improvements, and DRL. The review papers [5,26] explored a variety of topics, from conventional power system resilience methods to the metrics and assessment techniques for power system resilience. The review paper [27] comprehensively discussed machine learning strategies and their applications in conserving energy and managing it effectively, emphasizing their efficiency in addressing various decision and management challenges. Reference [28] discussed the transformation of power systems into cyber-physical systems (CPSs) and the unique resilience challenges posed by CPSs. The paper also highlighted the differences between conventional power systems and cyber-physical power systems (CPPSs), delving further into the realm of cyber-physical disturbances, resilience techniques for CPPSs confronting natural hazards and cyber threats, and the intriguing dimension of leveraging social behaviors to enhance CPPS resilience. Additionally, the review paper [29] scrutinized RL techniques in the context of power and energy systems, covering RL concepts, algorithmic diversity, and real-world applications, while considering the future course of RL within these domains. Reference [30] explored the utility of RL in navigating the intricacies of energy systems, offering an insightful classification of RL applications within the energy landscape and highlighting its potential to tackle rising system complexities, even as it struggles with utilization and benchmarking challenges. Lastly, the review paper [31] unraveled the expansive domain of DRL in power systems, highlighting foundational concepts, modeling intricacies, algorithmic diversity, and contemporary advancements. However, amid this wealth of knowledge, a significant gap exists: the application of DRL to enhance resilience in power and energy systems. Notably, previous works have primarily focused on showcasing the capabilities and applications of DRL in these domains, often overlooking the need to provide a comprehensive overview of the existing limitations, challenges, and future avenues. Given the increasing reliance

on DRL for resilience enhancement within these domains, this review paper emerges as a pivotal and missing link, illuminating the transformative potential of DRL in steering the trajectory of resilient power and energy systems.

In this article, an overview of the current progress in applying DRL to enhance the resilience of power and energy systems is presented. The foundations of different categories of DRL methods, including value-based, policy-based, and actor–critic methods, are laid out to aid in the understanding of their applications. Subsequently, an exploration of DRL applications within various domains of resilient power and energy systems, including dynamic response, recovery and restoration, energy management and control, communications and cybersecurity, and resilience planning and metrics development, is undertaken, with a focus on outlining the detailed methodologies and contributions of these studies. This in-depth analysis of DRL applications is followed by a discussion of the challenges and limitations associated with incorporating DRL into resilient power and energy systems. These revelations shed light on the practical limitations and possible consequences that researchers and practitioners need to take into account. Finally, the future is examined and a roadmap of potential research possibilities in this area is provided. Novel discoveries and solutions are eagerly anticipated in this field of research.

The remainder of this review article is organized as follows: Section 2 provides an introduction to DRL and its various categories. Section 3 offers a detailed exploration of DRL applications across various dimensions of resilient power and energy systems. Section 4 examines the existing challenges, limitations, and opportunities for future research in DRL applications for power and energy system resilience. Finally, Section 5 summarizes the key findings and highlights the transformative potential of DRL in ensuring the resilience of critical power and energy infrastructure.

2. Deep Reinforcement Learning Foundations

In the investigation of DRL and its crucial role in enhancing the resilience of the power and energy systems, deep reinforcement learning foundations are regarded as fundamental. They are indispensable in providing a solid foundation before setting out on this adventure by exploring the fundamental ideas that support DRL and its variety of algorithms. This section will outline the fundamentals of RL, the revolutionary potential of deep learning, and the powerful synergy that develops as these two fields converge to create DRL. A foundation will also be laid for a thorough understanding of different DRL algorithms' usefulness in the context of power and energy system resilience.

2.1. Reinforcement Learning (RL)

RL is characterized as a notable and dynamic machine learning paradigm in which an agent interacts continuously with its designated environment. The typical RL framework, depicted in Figure 2, consists of two main components: an artificial intelligence (AI) agent and the environment, engaging in reciprocal interactions until the agent achieves a learned state. In this context, it is typical to represent the environment as a Markov decision process (MDP), a common framework employed by numerous RL algorithms in this field by leveraging dynamic programming, as discussed in recent work by Xiang et al. [32]. A fundamental differentiator between classical dynamic programming techniques and RL algorithms is that the latter do not require precise knowledge of a mathematical model of the MDP. Instead, they focus on addressing large MDPs, where exact methods become impractical.

Within the realm of RL, the agent assumes the role of an autonomous decision-maker, persistently seeking choices that maximize a cumulative reward signal over an extended timeframe while navigating the intricacies of the environment [33]. This iterative process unfolds over discrete time steps, showcasing the agent's decision-making capabilities. At each time step, the agent observes the current state of the environment, assimilating vital information that guides its future actions. Empowered with this knowledge, the agent

makes deliberate decisions that impact not only its immediate actions but also have the potential to influence subsequent states and rewards within the environment.

Fundamental to the entire RL framework is the core objective of establishing optimal policies, serving as clearly defined action-selection strategies. These optimal policies are central in the RL journey, directing the agent towards decisions that hold the promise of the highest expected cumulative reward [34]. The agent navigates its course through the environment using this expected cumulative reward, which is frequently expressed as the expected return. In RL, learning is an ongoing process of exploration and improvement that uses a trial-and-error approach. The agent's actions are carefully chosen to best serve its long-term goals and are not random. Through a continuous feedback loop with the environment, this strategic decision making continuously evolves. The environment provides feedback to the agent after each action in the form of rewards or penalties. These evaluative indications give the agent essential information about the effects of its earlier actions, allowing it to gradually modify and improve its decision-making approach over time [35]. The underlying framework on which the RL paradigm is built is this complex interaction between the agent, environment, and the reward signals.

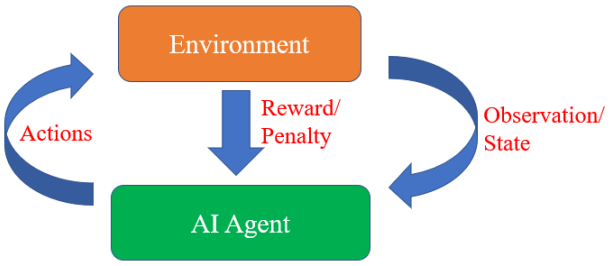

Figure 2. Typical framework of reinforcement learning.

2.2. Deep Learning

Deep learning, a subset of machine learning, makes use of deep neural networks, which are artificial neural networks with several layers. In order to provide output predictions, these networks methodically process input data through a series of complex transformations, forming an interconnected web of layers made up of neurons. The field of deep learning has produced amazing results in a variety of fields, including speech recognition, computer vision, natural language processing, and power systems. A transformative era and a significant advancement in the state of the art in a variety of tasks, including object detection, speech recognition, language translation, and power systems modeling, have been brought in by the emergence of deep learning [36].

One of the most compelling attributes of deep learning lies in its innate ability to automatically unearth concise, low-dimensional representations, commonly referred to as features, from high-dimensional datasets—ranging from images to text and audio. By incorporating inductive biases into the architectural design of neural networks, particularly through the concept of hierarchical representations, practitioners in the realm of machine learning have made significant strides in combating the notorious curse of dimensionality [37].

Furthermore, the impact of deep learning extends its reach into the domain of RL, catalyzing a distinct field known as DRL [38]. This fusion of deep learning algorithms with RL techniques has redefined the landscape of RL, unlocking new avenues for solving complex decision-making problems with unprecedented capabilities. The symbiotic relationship between deep learning and RL has ushered in a realm of possibilities for enhancing power and energy system resilience, where the exploitation of intricate patterns and representations in data holds substantial promise for addressing intricate challenges within the domain.

2.3. Deep Reinforcement Learning (DRL)

As outlined earlier, DRL marks the convergence of two influential domains: RL and deep learning. Within the realm of DRL, the utilization of neural networks, often extending into deep neural networks, plays a pivotal role in approximating the agent's policy or value functions [39]. This fusion empowers DRL with the ability to tackle complex decision-making tasks within high-dimensional state spaces, extending its applicability to a wide array of intricate challenges.

The foundation of deep learning predominantly rests on the capabilities of multilayer neural networks, where neurons serve as the fundamental building blocks [40]. The perceptron [41], one of the earliest neural network prototypes, initially surfaced as a single-layer neural network devoid of hidden layers. Its competence, however, was limited to straightforward linear classification tasks, rendering it incapable of resolving complex problems such as the XOR problem [42]. The ascent of multilayer perceptrons, characterized by an increased number of neurons and layers, brought forth remarkable nonlinear approximation capabilities. Hornik et al. [43] definitively established that multilayer perceptrons could approximate any nonlinear function, further solidifying the potential of deep learning in shaping the landscape of machine learning.

Two major success stories have emerged in the developing field of DRL, each of which marks a paradigm shift. The first, which served as the catalyst for the DRL revolution, was the creation of an algorithm capable of learning a wide variety of Atari 2600 video games at a superhuman level, straight from the raw image pixels [44]. By demonstrating that RL agents may be trained successfully using only a reward signal, even when faced with unprocessed, high-dimensional observations, this innovation resolved the long-standing instability challenges related to function approximation techniques in RL.

The development of AlphaGo, a hybrid DRL system that defeated a human world champion in the challenging game of Go, was the second significant accomplishment [45]. This success is comparable to IBM's victory over Deep Blue in the chess tournament two decades earlier [46]. Contrary to conventional rule-based chess systems, AlphaGo utilized the strength of neural networks trained using both supervised learning and RL, along with a standard heuristic search approach [38]. These outstanding accomplishments demonstrated DRL's excellent capacity for policy learning and optimization, enabling it to broaden its scope to address difficult real-world challenges for power and energy system resilience.

In the context of power and energy systems, a typical framework for training a DRL agent is illustrated in Figure 3. The DRL agent typically comprises multilayer neural networks, with an internal mechanism for continually updating the neural network weights. The environment generally includes a power system model and a reward generator, as depicted in the figure. The reward generator may employ system data and states obtained from the power system model to incentivize or penalize the DRL agent using rewards or penalties. The DRL agent takes the state from the power system model and provides actions to the power system model, in return for which it receives rewards or penalties. The figure represents a generic training framework for a DRL agent, and specific cases may involve some variations. It is important to note that during the implementation of the trained model, an actual power system may be used instead of a power system model.

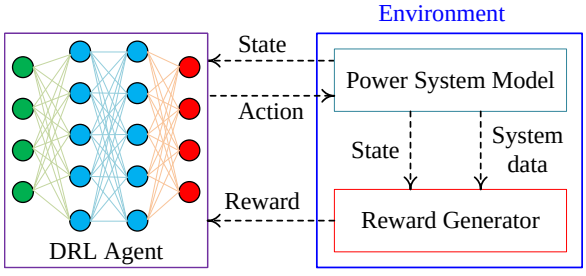

Figure 3. DRL training framework.

2.4. DRL Methods

In the realm of DRL, methodologies can be broadly classified into three primary categories: value-based, policy-based, and actor–critic methods, as categorized in [32] and as shown in Figure 4. Each of these categories will be briefly introduced in this subsection.

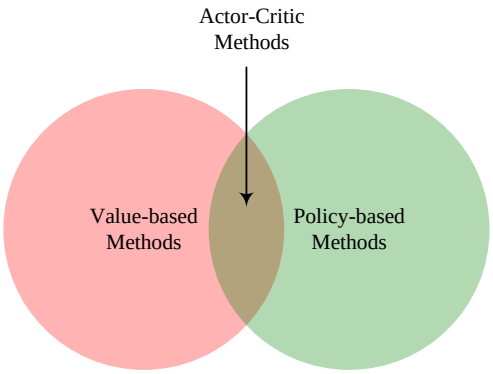

Figure 4. Different categories of DRL methods.

2.4.1. Value-Based Methods

Value-based DRL is regarded as a fundamental class of DRL methodologies, where the emphasis is placed on the representation of the value function and the determination of the optimal value function [39]. Within this category, the core objective is to capture and model the value function, a pivotal component that significantly influences the agent's decision-making processes.

The value function in value-based DRL serves as a critical guiding force for the agent. It provides insights into the expected cumulative rewards associated with taking various actions in specific states, aiding the agent in making informed choices. This methodological approach is characterized by its ability to approximate and optimize the value function, enabling the agent to navigate complex decision spaces effectively. Table 1 outlines merits and drawbacks of two popular value-based methods, namely, Q-learning and SARSA. These methods are discussed in detail below.

Table 1. Merits and drawbacks of various value-based methods.

Algorithm/Method	Merits	Drawbacks
Q-Learning	• Estimates the quality of actions through Q-values. • Aims to find optimal Q-values to maximize cumulative rewards. • Adaptable and efficient due to its off-policy nature. • Converges to an optimal policy that maximizes cumulative rewards.	• High variance and might not converge efficiently. • Often requires long training periods to estimate Q-values accurately. • Sensitive to the choice of hyperparameters, like the learning rate.
SARSA	• Employs an on-policy methodology to update Q-values. • Suitable for learning and improving the policy used during training. • Allows you to enhance the policy being used to interact with the environment.	• Might not find the optimal policy but instead optimizes the policy in use. • Could have slower convergence compared to Q-learning. • The on-policy nature makes it less efficient in some situations.

(a) Q-learning: Q-learning is a classic RL that focuses on estimating the quality, represented by Q-values, of taking a particular action in a given state within an environment [47]. Finding optimal Q-values is the main goal of Q-learning since they form an essential basis for directing the choice of actions in a way that maximizes cumulative rewards over time.

The Q-values for each state–action pair are updated iteratively by this algorithm depending on the knowledge gained from interacting with the environment. The predicted cumulative rewards that an agent can obtain by beginning in a particular state, executing

a particular action, and then implementing the best course of action are represented by Q-values. The strategy that results in the highest potential expected rewards is referred to as the optimal policy in this context.

By taking into account the immediate benefits gained from taking actions and the highest Q-value possible in the subsequent state, Q-learning iteratively improves its estimates of Q-values during the learning process. It accomplishes this utilizing a formula that strikes a balance between the knowledge that is currently known and the new information that is yet to be learned.

Q-learning is characterized by its off-policy nature, which allows it to learn from actions resulting from any policy, even if it randomly explores the environment. This characteristic adds to its adaptability and efficiency in a variety of applications.

The Q-value update function equation for Q-learning, also known as the Bellman equation, is typically expressed as follows [34]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot \left[r + \gamma \cdot \max_{a} Q(s',a) - Q(s,a) \right] \quad (1)$$

where $Q(s,a)$ represents the Q-value for a particular state–action pair (s,a), where s is the current state, and a is the action taken; α is the learning rate, which controls the extent to which the new information obtained from the update affects the Q-value; r stands for the immediate reward received after taking action a in state s; γ is the discount factor, representing how much importance is given to future rewards, which has a value between 0 and 1, with higher values indicating a greater emphasis on long-term rewards; s' represents the resulting state after taking action a in state s; and $\max_a Q(s',a)$ represents the maximum Q-value for all possible actions a in the new state s', which reflects the agent's estimate of the highest cumulative reward achievable from the new state onward.

The fundamental component of Q-learning is the update Equation (1), which enables the Q-values to repeatedly interact with the environment and iteratively converge toward their optimal values. By revising its Q-values in response to observed rewards and anticipated rewards from future states, the agent learns to make better decisions. Q-learning eventually reaches an optimal policy that optimizes the anticipated cumulative rewards.

(b) SARSA: An on-policy reinforcement learning technique referred to as SARSA, or "state–action–reward–state–action", shares similarities with Q-learning. It differs from Q-learning, though, in that it adopts an on-policy methodology. When updating the Q-values for a certain state–action combination, SARSA chooses a new action based on the existing policy and then updates the Q-values using the reward of the new action and the state that results from it [48]. In contrast, regardless of the policy being followed, Q-learning chooses the action with the highest anticipated reward for the subsequent state to update its Q-values. SARSA is ideally suited for situations where you wish to learn and enhance the policy being used to interact with the environment during training. This is because it has an on-policy capability.

The Q-value function update equation for SARSA is as follows [34]:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot \left[r + \gamma \cdot Q(s',a') - Q(s,a) \right] \quad (2)$$

where $Q(s,a)$ represents the Q-value for the current state–action pair (s,a); α is the learning rate, controlling the step size of the update; r is the immediate reward received after taking action a in state s; γ is the discount factor, determining the importance of future rewards; $Q(s',a')$ is the Q-value for the next state–action pair (s',a') after taking action a in state s.

By taking into account the immediate reward, the anticipated future rewards, and the learning rate, the updated Equation (2) determines the new Q-value for the current state–action combination. This reflects the SARSA algorithm's on-policy nature, as it uses the same policy to select the action a' in the next state s' for the update.

2.4.2. Policy-Based Methods

A policy-based method is an approach to RL/DRL where the agent learns a policy that directly maps states to actions. Unlike value-based DRL methods, which estimate the value of being in a particular state or taking a specific action, policy-based methods aim to find the optimal policy itself, which is a strategy for selecting actions in different states to maximize the expected cumulative reward. Table 2 outlines the merits and drawbacks of some popular policy-based methods. These methods are discussed in detail below.

Table 2. Merits and drawbacks of various policy-based methods.

Algorithm/Method	Merits	Drawbacks
Vanilla Policy Gradient (VPG)	• Focuses on maximizing expected cumulative rewards. • Adaptable due to its relative advantage and policy updates. • Theoretically sound and widely applicable. • Can handle high-dimensional action spaces.	• High variance in gradients, which can lead to slow convergence. • Requires careful tuning of hyperparameters like learning rates. • Limited sample efficiency in some environments.
Trust Region Policy Optimization (TRPO)	• Maintains policy consistency and prevents abrupt changes. • Uses a trust region constraint for safe policy updates. • Effective in continuous action spaces. • Stabilizes learning and avoids policy collapse.	• Computational demands due to solving constrained optimization problems. • May require complex optimization techniques (e.g., conjugate gradient). • Less straightforward to implement compared to simpler algorithms.
Proximal Policy Optimization (PPO)	• Provides a stable and sample-efficient alternative to TRPO. • Uses a clipped objective function to prevent policy divergence. • Relatively easy to implement and scale to various environments. • Suitable for both discrete and continuous action spaces.	• Can converge to sub-optimal solutions due to the clipping. • Requires careful tuning of hyperparameters, including the clipping range. • Might need a large number of samples for complex tasks.

(a) Vanilla policy gradient (VPG): The VPG algorithm is an RL method used to optimize policies in order to maximize the expected cumulative reward [48]. It focuses on comprehending the relative advantage of performing a particular action in a particular state versus choosing an action at random under the current policy. This relative advantage is captured by the advantage function, denoted as $A_\pi(s,a)$, and is defined as the difference between the action-value function $Q_\pi(s,a)$ and the state-value function $V_\pi(s)$, as shown below [48]:

$$A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s) \tag{3}$$

where $A_\pi(s,a)$ is the advantage of taking action a in state s under policy π; $Q_\pi(s,a)$ is the action-value function, representing the expected cumulative reward of taking action a in state s and following policy π thereafter; $V_\pi(s)$ is the state-value function, representing the expected cumulative reward when starting in state s and following policy π thereafter.

The VPG algorithm aims to find the policy parameters θ that maximize the expected return. This is accomplished through policy adjustments that increase the expected return. The following equation demonstrates how to calculate the policy gradient as the predicted value of the gradient of the policy with regard to its parameters, weighted by the advantage function [48]:

$$\nabla J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) A_\pi(s_t,a_t)\right] \tag{4}$$

where $\nabla J(\pi_\theta)$ is the policy gradient, representing the direction in which policy parameters should be updated; τ represents a trajectory generated by the policy π_θ, consisting of states s_t and actions a_t for each time step t; $\nabla_\theta \log \pi_\theta(a_t|s_t)$ is the gradient of the log-probability

of selecting action a_t in state s_t with respect to the policy parameters θ; and $A_\pi(s_t, a_t)$ is the advantage of taking action a_t in state s_t under policy π_θ.

The policy parameters θ are updated iteratively using this policy gradient, typically through gradient ascent, to improve the policy's performance over time. The VPG algorithm is one of the foundational methods in policy gradient RL, with variations like the REINFORCE algorithm that build upon this concept [49].

(b) Trust region policy optimization (TRPO): TRPO is a policy gradient algorithm introduced by Schulman et al. [50] in 2015. TRPO is renowned for its role in preserving the consistency of agent training for DRL. Its main goal is to stop excessive policy modifications that could otherwise cause performance to collapse. TRPO runs in a predetermined trust region, which denotes a particular parameter space where updating policies is thought to be safe. Its main goal is to find policy changes that uphold this trust region constraint while maximizing predicted rewards. In order to accomplish this, TRPO uses an iterative strategy that includes local approximations and policy adjustments made in accordance with the boundaries of the trust region. TRPO adds a penalty term based on the Kullback–Leibler (KL) divergence [51] between the new and old policies to make sure that the updated policy continues to stay close to the previous one.

The core idea of TRPO can be summarized in the following optimization problem with a maximization-type objective function (5), given below:

$$J(\pi) = \mathbb{E}_{s,a\sim\pi}\left[\frac{\pi(a|s)}{\pi_{\text{old}}(a|s)} A^{\pi_{\text{old}}}(s,a)\right] \tag{5}$$

where $J(\pi)$ represents the expected return under the new policy π; s and a are states and actions sampled from the policy; $A^{\pi_{\text{old}}}(s,a)$ is the advantage function computed with respect to the old policy π_{old}; and $\pi_{\text{old}}(a|s)$ is the probability of taking action a in state s according to the old policy.

The optimization problem is subject to a KL-divergence constraint given by (6), which ensures that the new policy is not too different from the old policy:

$$\mathbb{E}_{s\sim\pi_{\text{old}}}[D_{KL}(\pi_{\text{old}}(\cdot|s) \,||\, \pi(\cdot|s))] \leq \delta \tag{6}$$

where D_{KL} is the KL divergence between the old and new policies; and δ is the trust region radius, which limits how far the new policy can deviate from the old policy.

TRPO then solves this constrained optimization problem using various optimization techniques, such as conjugate gradient or natural gradient methods, to update the policy in a way that maximizes expected returns while respecting the trust region constraint. This ensures that policy updates are stable and do not lead to drastic performance degradation.

(c) Proximal policy optimization (PPO): PPO is an RL algorithm that was developed as a more straightforward substitute for TRPO. In terms of stability and sample effectiveness, PPO and TRPO are comparable, although PPO is easier to execute. Unlike TRPO, PPO uses a particular clipping method in its goal function rather than a KL-divergence restriction to make sure that the new policy stays close to the old policy [52].

The key features of PPO include its simplicity and improved sample efficiency, making it a popular choice for training RL agents. The objective function in PPO is defined as follows:

$$L(\theta) = \mathbb{E}\big[\min\big(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)\hat{A}\big)\big] \tag{7}$$

where $L(\theta)$ represents the objective function to be maximized with respect to the policy's parameters θ. The term $r(\theta)$ is the ratio of probabilities of taking actions under the new policy π_θ to the old policy $\pi_{\theta_{\text{old}}}$:

$$r(\theta) = \frac{\pi_\theta(a|s)}{\pi_{\theta_{\text{old}}}(a|s)} \tag{8}$$

where $\pi_\theta(a|s)$ is the probability of taking action a in state s under the new policy; and $\pi_{\theta_{\text{old}}}(a|s)$ is the probability of taking action a in state s under the old policy.

The term $\hat{A}$ represents the advantage function, which estimates the advantage of taking action a in state s under the current policy:

$$\hat{A} = Q(s, a) - V(s) \tag{9}$$

where $Q(s, a)$ is the state-action value function (Q-function), representing the expected return of taking action a in state s; and $V(s)$ is the state-value function, representing the expected return of being in state s under the current policy.

The clip function is used to clip the value of $r(\theta)$ within the range $[1 - \epsilon, 1 + \epsilon]$, where ϵ is a hyperparameter that controls the clipping range. Clipping helps to prevent the policy update from moving too far from the old policy, ensuring more stable updates.

Equation (7) illustrates how the objective function seeks to maximize the expected value of the minimum between two terms. The terms are the product of the advantage estimate and $r(\theta)$, and the clipped form of the same product. This objective promotes policy revisions that enhance performance without significantly deviating from the previous policy.

To maximize the objective function and adjust the policy parameters, PPO often uses stochastic gradient ascent techniques like Adam or RMSProp. PPO is a popular choice for RL tasks because it produces consistent and effective policy updates by iteratively optimizing the policy using this clipped objective function.

2.4.3. Actor–Critic Methods

Actor–critic is a popular RL method that combines aspects of both value-based and policy-based RL approaches [53]. In the actor–critic framework, an RL agent consists of two primary components: actor and critic. The actor is responsible for selecting actions in the environment. In regard to a particular state, it learns a policy that specifies the probability distribution over possible actions. The actor's job is to investigate and choose the best course of action to maximize anticipated rewards. The critic, however, assesses the actor's behavior. It gains knowledge of the value function, which calculates the potential returns from a given state or state–action pair. By evaluating the effectiveness of the actor's performance, the critic offers the actor feedback.

Utilizing the value judgments of the critic to inform and enhance the actor's policy is the core idea behind actor–critic methods. Gradient ascent is a common technique for accomplishing this, in which the actor modifies its policy in a way that raises the expected return as determined by the critic.

The actor–critic architecture has several advantages, some of which are listed below:

- *Advantage estimation*: Actor–critic methods are able to estimate the advantage of performing a specific action in a specific state by adding the critic's value function. The actor can concentrate on activities that are more likely to result in greater rewards with the aid of this advantage estimate.
- *Stability*: Actor–critic methods often exhibit more stable learning compared to pure policy-based or value-based methods. The critic's value estimates provide a stable baseline for updating the actor's policy.
- *Sample efficiency*: Actor–critic methods can be more sample-efficient than pure policy-based methods because they make use of value estimates to guide learning.

Actor–critic algorithms come in various forms and can be implemented using different techniques, including deep neural networks. Deep deterministic policy gradient (DDPG), twin delayed deep deterministic policy gradient (TD3), and soft actor–critic (SAC) are popular actor–critic algorithms that have achieved success in both discrete and continuous action spaces. Table 3 outlines the merits and drawbacks of these methods. The details of these methods are discussed below.

Table 3. Merits and drawbacks of various actor–critic methods.

Algorithm/Method	Merits	Drawbacks
Deep Deterministic Policy Gradient (DDPG)	• Handles continuous action spaces effectively. • Combines actor and critic networks for policy learning. • Uses target networks for training stability. • Successfully applied in robotic control tasks.	• Requires careful hyperparameter tuning. • Sensitive to the choice of neural network architectures. • Can have high sample complexity in some scenarios.
Twin Delayed Deep Deterministic Policy Gradients (TD3)	• Addresses Q-value overestimation issues. • Improves training stability with twin critics. • Uses a "twin delay" approach for policy updates. • Reduces overfitting and enhances reliability.	• Increased computational complexity due to twin critics. • Complexity in maintaining the twin delay mechanism. • Sensitivity to hyperparameter choices.
Soft Actor–Critic (SAC)	• Handles continuous action spaces with stochastic policies. • Promotes exploration through entropy regularization. • Achieves remarkable success in various RL tasks. • Efficiently combines actor and critic networks.	• Complexity in optimizing the temperature parameter. • Requires significant computational resources. • Tuning of hyperparameters, especially entropy-related ones.

(a) Deep deterministic policy gradient (DDPG): DDPG is an RL algorithm designed for solving tasks with continuous action spaces [54]. By extending the actor–critic design to deep neural networks, it makes it possible to simultaneously learn a policy (the actor) and a value function (the critic). A neural network acts as the actor, taking the current state as input and generating a continuous action. It gains knowledge of a deterministic strategy that directly links states and actions. In other words, it calculates the best course of action in a particular situation. As in conventional actor–critic systems, the critic, on the other hand, is a different neural network that accepts a state–action pair as input and calculates the expected cumulative reward (Q-value) linked to performing that action in the given state. The critic's job is to offer feedback on the quality of actions chosen by the actor.

The critic network is trained to minimize the mean squared error (MSE) between its predicted Q-values and the target Q-values. The target Q-values are computed using the Bellman equation and a target network to stabilize training:

$$L(\theta_{\text{crc}}) = \mathbb{E}[(Q(s,a|\theta_{\text{crc}}) - (r + \gamma Q(s', \mu(s'|\theta_{\text{tgt\_act}})|\theta_{\text{tgt\_crc}}))^2] \quad (10)$$

where $Q(s,a|\theta_{\text{crc}})$ is the Q-value predicted by the critic network; r is the immediate reward received after taking action a in state s; γ is the discount factor; and $Q(s', \mu(s'|\theta_{\text{tgt\_act}})|\theta_{\text{tgt\_crc}})$ is the target Q-value estimated using the target actor network to select the next action.

The actor network is updated based on the deterministic policy gradient. The actor aims to maximize the expected cumulative reward with respect to its parameters:

$$\nabla_{\theta_{\text{act}}} J \approx \mathbb{E}[\nabla_a Q(s,a|\theta_{\text{crc}})|_{s=s_t, a=\mu(s_t|\theta_{\text{act}})} \nabla_{\theta_{\text{act}}} \mu(s|\theta_{\text{act}})|_{s=s_t}] \quad (11)$$

where θ_{act} are the actor's parameters; θ_{crc} are the critic's parameters; J is the expected cumulative reward; $\mu(s|\theta_{\text{act}})$ is the actor's policy (action) in state s; and $Q(s,a|\theta_{\text{crc}})$ is the Q-value predicted by the critic network.

To stabilize training, DDPG uses target networks (target actor and target critic). To provide target Q-values and target actions, these target networks are soft-copied from the main networks on a regular basis. A broad variety of robotic control and continuous control problems in RL have been successfully implemented using DDPG, which is renowned for its capacity to handle continuous action spaces. It is a successful method for learning complex control policies because it strikes a balance between exploration and exploitation in continuous action domains.

(b) Twin delayed deep deterministic policy gradients (TD3): TD3 [55] is an advanced RL algorithm that builds upon the foundation of the DDPG algorithm. TD3 was developed

to address some of the key challenges associated with training DRL agents, particularly in tasks with continuous action spaces. The critic network's tendency to overestimate Q-values in DDPG and other comparable algorithms is a key problem that can cause training instability and poor performance. By using twin critics—basically, two distinct Q-value estimation networks—TD3 tackles this problem. It takes the minimum of the Q-values provided by these twin critics as the target value during the learning process. This helps in reducing the overestimation bias and leads to more accurate value estimates.

The addition of twin delays for updating the actor (policy) and target networks is another significant aspect of TD3. TD3 updates the actor network less frequently, typically after every two updates to the critic network, in contrast to conventional algorithms that update both the actor and critic networks simultaneously. This "twin delay" approach improves training stability and reduces overfitting problems, making TD3 more reliable in real-world applications. During the learning process, TD3 also adds noise to the target action and target policy. By regularizing the learning process, this noise injection makes it harder for the agent's policy to take advantage of Q-value estimation errors. Overall, TD3's advances work together to make DRL training more efficient and reliable, especially in continuous action space settings where accurate action selection is essential.

(c) Soft actor–critic (SAC): SAC [56] is an advanced RL algorithm designed for tasks with continuous action spaces. In SAC, the actor–critic framework is improved, and entropy regularization is incorporated to promote exploration and improve stochastic policies. SAC is renowned for its ability to handle challenging continuous control problems and produce improved action space exploration.

Being an actor–critic type of algorithm, SAC consists of actor (policy network) and critic (Q-value network). The actor network, denoted as $\pi(a|s)$, is a stochastic policy that maps states s to probability distributions over actions a. In contrast to deterministic policies, SAC represents policies using a probabilistic method. It models the policy as a Gaussian distribution with mean $\mu(s)$ and standard deviation $\sigma(s)$. The policy network's output represents the parameters of this distribution. Similarly, the critic network, denoted as $Q(s, a)$, estimates the expected cumulative reward (Q-value) associated with taking action a in state s. Like in traditional actor–critic methods, the critic's role is to provide feedback on the quality of actions selected by the actor. To promote exploration, SAC adds an entropy term to the objective function. The entropy regularization term is defined as $H(\pi(s))$, where H represents the entropy of the policy distribution. The objective function is a combination of the expected cumulative reward (Q-value) and entropy:

$$J(\theta) = \mathbb{E}_{(s,a)\sim\rho_\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) + \alpha H(\pi(s))\right] \tag{12}$$

where θ represents the parameters of both the actor and critic networks; ρ_π is the state-action distribution induced by the policy π; γ is the discount factor; $r(s_t, a_t)$ is the immediate reward received after taking action a_t in state s_t; and α is a temperature parameter that determines the tradeoff between exploration (entropy) and exploitation (reward).

The actor network is updated to maximize the combined objective function, which includes the expected cumulative reward and the entropy term. The actor aims to find the policy parameters that maximize the expected reward while also maximizing entropy:

$$\nabla_{\theta_{\text{actor}}} J(\theta_{\text{actor}}) = \mathbb{E}_{s\sim\rho_\pi}[\nabla_{\theta_{\text{actor}}} \log \pi(a|s, \theta_{\text{actor}}) Q(s, a|\theta_{\text{critic}})] + \alpha \nabla_{\theta_{\text{actor}}} H(\pi(s|\theta_{\text{actor}})) \tag{13}$$

The critic network is updated to minimize the mean squared error (MSE) between its predicted Q-values and the target Q-values. The target Q-values are computed using the Bellman equation:

$$L(\theta_{\text{critic}}) = \mathbb{E}[(Q(s, a|\theta_{\text{critic}}) - (r + \gamma(1 - \text{done}) Q(s', \pi(s'|\theta_{\text{actor}})|\theta_{\text{critic}}))^2] \tag{14}$$

where s' is the next state; and done is an indicator variable that is 1 if the episode terminates at the next step and 0 otherwise.

The temperature parameter α in (12) and (13) is updated through optimization to find the optimal tradeoff between exploration and exploitation.

SAC is renowned for its capability to manage continuous action areas with efficiency while fostering exploration via entropy regularization. It has achieved remarkable success in various tasks including robotic control [57] and Volt-VAR optimization [58], making it a valuable algorithm in the field of RL.

3. Deep Reinforcement Learning Applications in Different Aspects of Resilient Power and Energy Systems

This section presents an in-depth exploration of the multifaceted applications of DRL within various critical aspects of resilient power and energy systems. Power and energy system resilience refers to the capacity of a power and energy infrastructure to endure, absorb, and promptly recover from various disruptions, including natural disasters and man-made events, while maintaining the continuity and reliability of power supply to end consumers [59]. This concept acknowledges the increasing challenges posed by extreme weather events, cyber-attacks, climate change, and the need for adaptive responses in the power sector. Evaluating the resilience of a complex system, particularly in the context of power and energy systems, necessitates a comprehensive and systematic approach. The Disturbance and Impact Resilience Evaluation (DIRE) methodology offers precisely such a framework for assessing and enhancing resilience [60]. In Figure 5, the various stages within the DIRE approach are illustrated, each of which plays a critical role in resilience assessment and evaluation. The DIRE framework consists of five distinct stages, namely, reconnaissance (recon), resist, respond, recover, and restore. These stages are integral in assessing a system's ability to endure and adapt to disruptions. DRL emerges as a powerful tool that finds applications across all of these resilience stages. DRL offers the capacity to adapt, optimize, and enhance system behavior in response to evolving conditions and disturbances, making it a valuable asset in the pursuit of resilience.

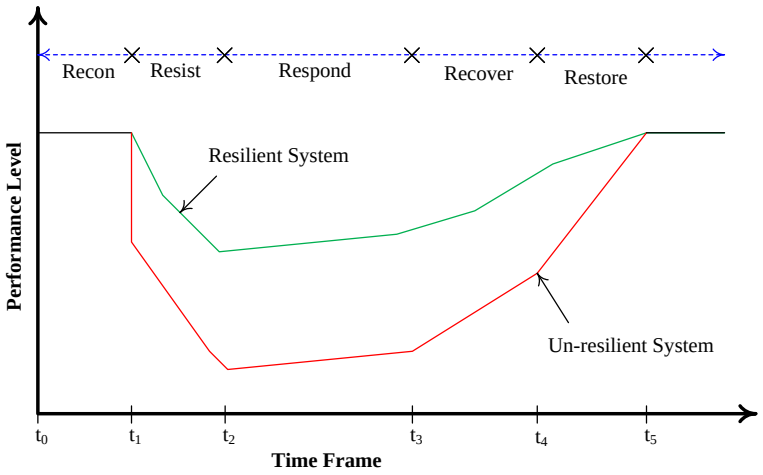

Figure 5. The DIRE curve showing different stages of resilience.

Figure 5 not only depicts the stages but also showcases typical DIRE curves for both resilient and unresilient systems. These curves offer a visual representation of how system performance evolves over time in the face of extreme events and disturbances. It is clear from the figure that after an extreme event occurs, at time t_1, the performance level of the system undergoes a deterioration. This decline in performance persists throughout the "resist" stage until reaching time t_2. However, the "respond" stage, which extends to t_3, marks the onset of a slow but steady performance improvement. Subsequently, as the

system enters the "recover" and "restore" stages, its performance continues to improve, ultimately returning to pre-disturbance levels.

Power and energy system resilience encompasses a spectrum of research areas and methodologies, including the development of resilience metrics, resilience planning, and operational resilience enhancement, each addressing critical aspects of ensuring the robustness of power and energy systems. Figure 6 shows five different aspect categories of resilient power and energy systems that are outlined in this section. Understanding these aspects of resilient power and energy systems is critical for devising comprehensive strategies to ensure the reliability and robustness of power and energy infrastructure in the face of evolving challenges and disruptions. Each aspect contributes to a holistic approach aimed at enhancing the resilience of power and energy systems and safeguarding the uninterrupted supply of electricity to end consumers.

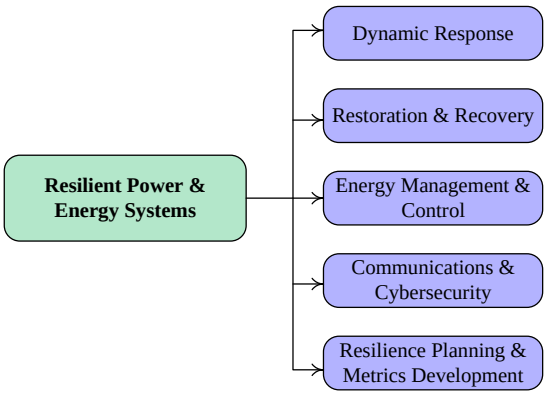

Figure 6. Different aspects of resilient power and energy systems.

3.1. Dynamic Response

Dynamic response consists of adaptive measures and strategies aimed at addressing and mitigating the consequences of unforeseen and critical events or disasters, whether natural or human-made. These events may include hurricanes, earthquakes, floods, wildfires, cyber-attacks, industrial accidents, or acts of terrorism. Dynamic response can operate proactively, linking situational awareness with resilience enhancement and ensuring effective and efficient responses in both preventive and emergency contexts [61]. The primary objectives of emergency response in the context of power and energy systems resilience include ensuring the integrity and functionality of critical infrastructure, minimizing downtime, and restoring operations swiftly and effectively. Table 4 presents a summary of papers on DRL applications in the dynamic response aspect of resilient power and energy systems.

In [62], a new model-free online dynamic multi-microgrid formation (MMGF) scheme for enhancing power system resilience was introduced. This approach formulated the dynamic MMGF challenge as a Markov decision process and tailored a comprehensive DRL framework for microgrids capable of changing their topologies. To address the complexity of operating network switches, a search space reduction technique based on spanning forest was introduced, and an action-independent value function was implemented. Subsequently, an advanced approach utilizing a CNN-based double DQN was designed to enhance the learning capabilities beyond the original DQN approach. This DRL approach has been supposed to provide real-time computational support for online dynamic MMGF, effectively addressing long-term resilience enhancement challenges by adaptively forming microgrids in response to changing conditions. Validation of the proposed method was carried out using both a 7-bus system and the IEEE 123-bus system, demonstrating its robust learning ability, timely response to varying system conditions, and effective enhancement of resilience.

Table 4. Summary of papers on DRL applications in dynamic response.

Paper (Authors and Reference)	DRL Algorithm	Main Contributions
Zhao et al. [62]	CNN-based double DQN	MDP-based multi-microgrid formation for enhanced resilience using CNN-based double DQN.
Zhou et al. [63]	Q-learning, VPG, and PPO	Framework for rescheduling optimization in distribution systems, applicable in various network topologies.
Kamruzzaman et al. [64]	Hybrid soft actor–critic	Multi-agent voltage violation mitigation during windstorms, scalable and flexible control.
Chen et al. [65]	Deep recurrent Q-network (DRQN)	Active power correction control in large-scale power systems through cooperative stochastic games.
Abdelmalak et al. [66]	Actor–critic algorithm	Distribution network reconfiguration during extreme weather events with high action accuracy.
Kadir et al. [67]	DDPG	Proactive power grid control during wildfires, minimizing outages and improving emergency response.
Badakhshan et al. [68]	PPO	RL-based intentional islanding for voltage stability during outages.
Huang et al. [69]	DQN	Model-free microgrid formation for resilient distribution networks, addressing complex scenarios.
Liang et al. [70]	Safe reinforcement learning with combination of DQN and CVaR method	Resilient proactive scheduling for commercial buildings during extreme weather, optimizing comfort while conserving energy.

In [63], a model-free optimization framework rooted in DRL was introduced to optimize rescheduling strategies, thereby enhancing the responsiveness of resilient distribution systems. The proposed approach leveraged deep neural networks to extract and process extensive features from a complex and stochastic state space. Multivariate Gaussian distributions were employed to manage high-dimensional continuous control actions effectively. The study conducted comprehensive testing and comparison of the proposed framework against a basic RL method (i.e, Q-learning) and two distinct DRL algorithms (VPG and PPO) across various power system configurations, including the IEEE 9-bus, IEEE 39-bus, and IEEE 123-bus systems. The empirical results underscored the remarkable effectiveness of the proposed DRL-based framework in both radial and meshed topologies.

In [64], a data-driven multi-agent framework based on a DRL algorithm was developed to address power system resilience in emergency scenarios, focusing on voltage violations during windstorms. The framework utilized a hybrid soft actor–critic algorithm for both offline and online tasks related to shunt reactive power compensators' deployment and control. The multi-agent system learned from past experiences to determine optimal locations and sizes for shunts, mitigating voltage violations during line failures. Demonstrated on the IEEE 57-bus and IEEE 300-bus systems, the approach effectively improved power system resilience by enhancing the scalability of existing DRL-based methods, offering flexibility for various control problems, and ensuring a balance between exploration and exploitation. This work streamlined information sharing among actors and critics, reducing the computational burden for large-scale integrated power systems.

In [65], a distributed DRL framework was employed to address the active power correction control (APCC) problem in large-scale power systems. They established an APCC

model, including topology reconfiguration actions, and utilized a fully cooperative stochastic game to represent interactions among active power controllers. A model-free approach was adopted to find the Nash equilibrium for the game, with the discrete nature of the APCC problem addressed using the QMIX method. To meet practical application requirements for large-scale power systems, they proposed a structure involving centralized online training and offline distributed execution. The method's effectiveness was verified using an open-source platform with relevant scenarios and cases. This research demonstrated the potential of distributed DRL in managing complex power systems, opening avenues for future work to address control problems with high dimensionality and nonlinearity.

In [66], an RL-based method has been introduced for distribution network reconfiguration (DNR) aimed at improving the resilience of electric power supply systems. Resilience improvements often involve tackling computationally intensive and sometimes infeasible large-scale stochastic optimization challenges. The study leveraged the exceptional performance of RL techniques in power system control applications, particularly for real-time resilience-based scenarios. They developed a single-agent framework using an actor–critic algorithm (ACA) to determine the statuses of tie-switches in a distribution network impacted by extreme weather events. The approach reconfigured the feeder topology to minimize or eliminate load shedding, treating the problem as a discrete Markov decision process. Actions involved opening or closing specific tie-switches, with rewards computed to evaluate their practicality and advantages. Through iterative Markov processes, the ACA was trained under various failure scenarios and demonstrated on a 33-node distribution system, achieving action accuracy exceeding 93%.

In [67], the authors worked on addressing the proactive control problem within the context of resilient power and energy systems during extreme events, specifically wildfires. They formulated the problem as an MDP to minimize load outages, taking into account the dynamic nature of wildfire events. Their approach introduced an integrated testbed that combined a wildfire propagation simulator with a power-system simulator, enabling comprehensive evaluations. By leveraging DRL for power generation coordination, their approach aimed to provide intelligent proactive control for power grid operators. The results demonstrated its effectiveness in reducing load outages during extreme events, offering a valuable contribution to enhancing emergency response in power and energy systems.

In [68], the authors developed an RL-based model to facilitate intentional islanding in power systems, offering real-time switching control and adaptability to changing system conditions. They framed the intentional islanding process as an MDP and utilized RL to learn the optimal transmission switching policy. This approach incorporated a PSS/E model of the power transmission system, interfacing with the OpenAI Gym [71]. The primary goal was to create stable and self-sustainable islands while maintaining voltage stability and reducing power mismatches. They implemented a PPO algorithm with multilayer perceptrons for policy and value networks to control switch statuses. Their framework's effectiveness in grid self-recovery through intentional islanding was validated on a modified IEEE 39-bus test network using dynamic simulations, demonstrating its potential for online topology control in transmission networks during outages.

In [69], the authors worked on addressing the need for resilience in power grids during extreme weather events, focusing on microgrid formation for resilient distribution networks (RDNs) when the primary utility power is unavailable. They proposed a model-free RDN-oriented microgrid formation method based on DRL techniques. The approach treated microgrid formation as an MDP, considering intricate factors like unbalanced power flow analysis of the microgrid along with its operational limitations. They developed a simulation environment utilizing OpenAI Gym [71], facilitating the application of the DRL methodology. The DQN was employed to find the optimal configuration of microgrids. Furthermore, a framework based on an offline training and real-time implementation was developed. Thorough examination of numerical results was conducted, offering a model-free solution for microgrid formation with complex scenarios and fast application

efficiency. Key contributions included the MDP formulation, simulator-based environment, and the introduction of DRL for near-optimal solutions.

In [70], the authors worked on addressing the challenge of ensuring the resilience of commercial buildings (CBs) during high-impact, low-frequency extreme weather events. They introduced a resilient proactive scheduling strategy based on safe reinforcement learning (SRL) to optimize customer comfort levels while minimizing energy reserve costs. The strategy leveraged the correlation between various CB components equipped with demand response capabilities to maintain desired comfort levels with limited energy reserves. To handle uncertainties associated with extreme weather events, they developed an SRL algorithm by combining DQN and conditional-value-at-risk (CVaR) methods. This approach enabled proactive scheduling decisions that balanced exploration and exploitation, effectively mitigating the impact of extreme events during the learning process. Extensive simulations using real-world commercial campus data demonstrated the capability of the proposed framework to enhance both power and heating/cooling resilience, ensuring power balance and maintaining comfort levels in CBs subject to extreme weather conditions. The study presented a valuable contribution by integrating correlated demand response with proactive scheduling to support maximum comfort levels while conserving energy reserves in commercial buildings exposed to extreme weather events.

3.2. Recovery and Restoration

The phases of recovery and restoration are of utmost importance in the context of power and energy system resilience. These phases cover the plans and methods used to restore order to the power and energy system after a disruptive event. Recovery and restoration are intrinsically linked to dynamic response methods, with subsequent recovery efforts being made easier by the initial response phase.

Recovery entails a diverse strategy aimed at determining the degree of damage, stabilizing the power and energy system, and starting the restoration and repair operations. This step comprises a thorough assessment of the state of the power system, including the identification of crucial elements that might have been jeopardized during the incident [72]. Critical loads quickly regain access to power due to recovery mechanisms that prioritize the restoration of key services.

On the other hand, restoration concentrates on the systematic approach of returning the entire power and energy system to its pre-disruption state [73]. In this phase, damaged infrastructure is coordinately repaired and reconnected, system integrity is tested and verified, and non-essential services are gradually brought back online. The goal of restoration efforts is to bring the power and energy system back to full functionality so that it can efficiently meet consumer needs [74].

This subsection examines the application of DRL approaches to speed up recovery and restoration procedures in power and energy systems. In-depth discussion is provided regarding how DRL may improve prioritizing, resource allocation, and decision making during these crucial times, thereby improving the overall resilience and dependability of the power and energy system. Table 5 presents a summary of papers on DRL applications in the restoration and recovery aspect of resilient power and energy systems.

In [75], a DRL-based model was developed to efficiently restore distribution systems following major outages. The model employed a Monte Carlo tree search (MCTS) to expedite training and enhance decision making in scenarios with partial and asynchronous information. Through iterative exploration and exploitation, the model determined restoration actions, including load prioritization, resource allocation, and power system performance evaluation. A reward function incentivized actions that served critical loads, ensured system security, and minimized restoration efforts. The integration of a power flow simulation tool (OpenDSS) and MCTS improved training performance and scalability. While the paper focused on distribution systems' resilience, the DRL-based model demonstrated potential applicability to resilience challenges at both the transmission and distribution levels,

particularly in the context of extreme events and the increasing complexity of modern distribution grids with DERs and low observability.

Table 5. Summary of papers on DRL applications in restoration and recovery.

Paper (Authors and Reference)	DRL Algorithm	Main Contributions
Bedoya et al. [75]	Deep Q-learning	DRL model for efficient restoration of distribution systems after outages.
Hosseini et al. [76]	DDPG	Intelligent resilience controller (IRC) for real-time operational decisions during outages.
Du et al. [77]	A modified DDPG	Two-stage learning framework for DER scheduling in microgrids.
Gautam [78]	DQN	Distribution system resilience through planning and operation-based strategies.
Dehghani et al. [79]	Advantage actor–critic (A2C)	Planning framework for enhancing distribution system resilience against climatic events.
Li et al. [80]	Q-learning	Integrated recovery strategy for maximizing electricity supply during grid restoration.
Zhao et al. [81]	Graph RL using Q-learning	Optimization of distribution service restoration in power grids.
Wang et al. [82]	Double DQN	Real-time routing and scheduling for MESSs.
Qiu et al. [83]	Hierarchical multi-agent PPO	Decentralized repair crew dispatch in coupled networks.
Nie et al. [84]	DQN and DDPG	Optimization of islanded microgrid operation with limited resources.
Abdelmalak et al. [85]	Multi-agent DDPG	Efficient DER dispatch for operational resilience of islanded power systems.
Gautam et al. [86]	VPG	Critical load restoration within active distribution systems.
Gautam et al. [87]	DQN	Optimization of distribution network reconfiguration during extreme events.
Yao et al. [74]	TD3	Optimal scheduling of MESSs for distribution system resilience.
Zhang et al. [73]	VPG	Improved load restoration using RL in uncertain renewable energy scenarios.
Wang et al. [88]	Hierarchical multi-agent PPO	Hierarchical method for coordination of mobile power sources and repair crews to enhance microgrid resilience.

In [76], DRL was employed to create an intelligent resilience controller (IRC) capable of making rapid real-time operational decisions for dispatching distributed generation and energy storage units to restore power following sudden outages. The IRC was designed to learn the patterns associated with uncertain high-impact events, exemplified by a spatiotemporal hurricane impact analysis model, and used this knowledge to explore a wide range of actions in partially observable states of distribution grids during extensive outages. The distribution grid's operation under uncertainty was modeled as an MDP, and operator actions were rewarded based on operational costs. To address scalability challenges due to several DERs, the problem was reformulated as a sequential MDP. Implementation of the

proposed model on a hurricane-affected test distribution grid demonstrated its superiority in terms of reduced operation costs and nearly instantaneous decision making, showcasing its adaptability to hurricanes of varying intensities. This study introduced an intelligent approach to enhance the resilience of distribution grids during extreme weather conditions, utilizing DRL and a spatiotemporal hurricane model to inform real-time dispatch decisions for DERs.

In [77], the authors addressed the challenge of efficiently scheduling and dispatching DERs in islanded microgrids to support service restoration and enhance system resilience during utility grid outages. Conventional model-based methods for DER coordination often lack generalization and adaptability, relying on precise distribution network models. To overcome these limitations, the authors proposed a two-stage learning framework based on deep deterministic policy gradient from demonstrations (DDPGfD). During the initial training phase, imitation learning was employed to provide the control agent with expert experiences, ensuring a satisfactory starting level of performance. In the subsequent online training phase, techniques such as reward shaping, action clipping, and the inclusion of expert demonstrations were harnessed to support secure exploration and expedite the training procedure. The proposed approach, applied to the IEEE 123-node system, was shown to outperform representative model-based methods and the standard DDPG method, demonstrating both solution accuracy and increased computational efficiency. This study introduced a pioneering application of DRL from demonstrations for distribution service restoration, addressing the high-dimensional complexity of continuous action spaces and providing a valuable contribution to enhancing microgrid resilience.

In [78], the concerns of electric utilities, government agencies, and the public regarding the impacts of natural disasters and extreme weather events on distribution systems' security, reliability, and resilience was addressed through the use of movable energy resources (MERs). The study focused on MERs, which are flexible and dispatchable to power outage locations following disasters. The dissertation encompassed three interdependent tasks for enhancing distribution system resilience using MERs. The first task involved determining the optimal total size and number of MERs, employing graph theory and combinatorial techniques. The second task focused on pre-positioning MERs using cooperative game theory, based on weather forecasts and monitoring data. The final task was post-disaster routing of MERs, which was addressed using a DRL-based model. The research's overarching thesis was that a combination of planning and operation-based strategies for MERs would significantly enhance the resilience of electric distribution systems.

In [79], the authors focused on enhancing the long-term resilience of power distribution systems faced with extreme climatic events by employing grid-hardening strategies. To address the challenges of limited budgets and resources, a planning framework based on DRL was developed. The resilience enhancement problem was treated as an MDP, and an approach was developed, combining ranking strategies, neural networks, and RL. Unlike conventional methods that target resilience against single future hazards, this framework quantified life-cycle resilience, considering the possibility of multiple stochastic events during the system's lifetime. A temporal reliability model was introduced to account for gradual deterioration and hazard effects, particularly in the context of stochastic hurricane occurrences. The framework was applied to a substantial power distribution system with thousands of poles, and the results demonstrated a significant improvement in long-term resilience compared to existing strategies, including the National Electric Safety Code (NESC) strategy, with an enhancement of over 30% for a 100-year planning horizon. Additionally, the DRL-based approach provided optimal solutions for computationally challenging problems that were difficult to solve using the branch and bound (BB) algorithm, making it a promising approach for enhancing system resilience in the face of extreme events.

In [80], the authors addressed the vulnerability of power grids to extreme events and the critical need for efficient restoration strategies to bolster grid resilience. They proposed an integrated recovery strategy that aimed to maximize the total electricity supplied to loads during the recovery process, taking into account various time scales of restoration

methods. This strategy harmoniously combined the gradual component repair process with the swift restoration method of optimal power dispatch. To achieve this, they utilized the Q-learning algorithm to determine the sequence for repairing damaged components and updating the network topology. Additionally, linear optimization was employed to maximize power supply on a given network structure. Simulation results, based on testing the proposed method on the IEEE 14-bus and IEEE 39-bus systems, indicated its effectiveness in coordinating available resources and manpower to swiftly restore the power grid following extreme events. This research is supposed to offer valuable insights for power operators aiming to enhance grid resilience.

In [81], a distribution service restoration (DSR) algorithm was presented as a crucial component of resilient power systems, offering optimal coordination for enhanced restoration performance. Typically, model-based control methods were employed for this purpose, but they suffered from low scalability and required precise models. To overcome these challenges, the study introduced an approach based on graph-reinforcement learning (G-RL). G-RL integrated graph convolutional networks (GCNs) with DRL to address the intricacies of network restoration in power grids, capturing interactions among controllable devices. The scalability of the solution was ensured by treating DERs as agents within a multi-agent environment. This framework utilized latent features derived from graphical power networks, which were processed by GCN layers to guide network restoration decisions through RL. Comparative evaluations conducted on the IEEE 123-node and 8500-node test systems demonstrated the effectiveness and scalability of the proposed approach, offering a versatile solution for dynamic DSR problems in power systems.

In [82], the authors focused on addressing the real-time routing and scheduling challenges posed by multiple mobile energy storage systems (MESSs) within power and transportation networks, to enhance system resilience. Prior research primarily employed model-based optimization for MESS routing, which was time consuming and demanded extensive global network information, raising privacy concerns. Moreover, real-time control of MESSs was challenging due to system variability. To address these issues, the study introduced a model-free real-time multi-agent DRL approach. In this approach, parameterized double DQNs were employed to reconfigure the coordination of MESS scheduling, accommodating both continuous and discrete types of actions. The RL environment was represented by the interconnected transportation network and a linearized AC optimal power flow solver, which integrated uncertainties arising from factors such as renewable generation, line outage information, etc., into the learning process. Thorough numerical studies performed on 6-bus and 33-bus power networks confirmed the superior performance of the proposed approach over traditional model-based optimization methods, making it a valuable contribution to resilience enhancement.

In [83], the authors focused on addressing the challenges posed by extreme events on microgrids, which could lead to significant outages and restoration costs. Repair crews (RCs) are essential for system resilience due to their mobility and adaptability in both transportation and energy systems. Coordinating RC dispatch is complex, particularly in multi-energy microgrids with dynamic uncertainties. The paper tackled this by formulating the RC dispatch problem in an integrated power–gas transportation network as an MDP. A hierarchical MARL algorithm was introduced, featuring a two-level optimization problem with switching decisions being the higher-level optimization and routing and repairing decisions of RCs the lower-level optimization. An abstracted critic network was integrated to capture the system dynamics and stabilize the training performance while preserving privacy. Thorough numerical investigations on an integrated power–gas transportation network demonstrated the algorithm's superiority over traditional MARL methods based on various efficiency metrics. The approach's scalability was also validated on a larger 33-bus power and 15-bus gas network with an 18-node 27-edge transportation network.

In [84], the authors focused on enhancing the post-disaster resilience of a distribution system that experiences power supply interruptions due to extreme disasters, forcing it to operate as an islanded microgrid. To optimize the utilization of limited generation

resources and provide extended power supply to critical loads during the outage, a multi-agent DRL approach was developed. This method implemented dual control policies for managing energy storage and load shedding within the microgrid. The goal was to maximize the cumulative utility value of the microgrid over the duration of the power outage. A comprehensive simulation environment, constructed using OpenAI Gym and OpenDSS, facilitated testing and validation. The results demonstrated the adaptability of the proposed approach under various conditions, including different available generation resources and microgrid outage durations. The main contributions of this work included the development of a multi-agent DRL model for sequential decision making, the formulation of dual optimal control policies for source and load sides, and the creation of an RL environment for islanded microgrid operation, accounting for generation limitations, power flow, and microgrid uncertainties.

In [85], an RL-based approach was introduced to dispatch DERs and enhance the operational resilience of electric distribution systems following severe outage events. The escalating computational intricacies and the need for intricate modeling procedures in resilience-based enhancement strategies prompted the adoption of intelligent algorithms tailored for real-time control applications. The authors utilized a multi-agent DRL algorithm to efficiently dispatch DERs in the aftermath of extreme events, with the primary objective of providing a swift and effective control mechanism for improved resilient operation of islanded distribution power systems. The problem was formulated as an iterative MDP, encompassing system states, action spaces, and reward functions. Each agent was responsible for dispatching a specific DER and was trained to maximize its cumulative reward value. System states represented the system's topology and characteristics, actions denoted DER power-supply decisions, and rewards were computed based on the power balance mismatch for each agent. The proposed model was trained using various failure scenarios and demonstrated on a 33-node distribution system in islanded mode, illustrating its capability to dispatch DERs for resilience enhancement.

In [86], a DRL approach was proposed for post-disaster critical load restoration within active distribution systems, with the aim of forming microgrids using network reconfiguration to reduce the amount of curtailed critical loads. The power distribution networks (PDNs) were modeled using graph theory, and the best system configurations, involving microgrids, were determined by finding the optimal spanning forest while adhering to various distribution system-related constraints. In contrast to existing methods, which required repetitive calculations for every line outage case to determine the optimal spanning forest, the proposed methodology, once adequately trained, could rapidly identify the best spanning forest even when line outage cases varied. In situations involving multiple line failures, the DRL-based model established microgrids with DERs, minimizing the need for curtailing critical loads. The model underwent training using the REINFORCE algorithm, an RL technique that relies on the VPG method. The paper was concluded with a numerical investigation performed on a 33-node distribution system, illustrating the capability of the proposed methodology in restoring critical loads after a disaster.

In [87], a DRL-based approach aimed at optimizing the reconfiguration of PDNs to improve their resilience against extreme events was presented, with the primary objective of reducing the amount of curtailed critical loads. The PDN was depicted as a graph theoretic network, and the optimal system configuration was determined by seeking the optimal spanning forest while adhering to various distribution system operational constraints. Differing from traditional techniques that necessitate repeated calculations for each system operating state to find the optimal network configuration, the DRL-based approach, once adequately trained, exhibited the capability to rapidly identify the optimal or near-optimal configuration even in the face of changing system states. To minimize critical-load curtailment resulting from multiple line outages during extreme events, the proposed approach formed microgrids incorporating DERs. A DQN-based model was utilized. The paper was supported by a numerical analysis conducted on a 33-node distri-

bution test system, confirming the efficiency of the proposed methodology for enhancing PDN resilience through reconfiguration.

In [74], the authors explored the utilization of mobile energy storage systems (MESSs) to bolster distribution system resilience. They devised an MDP framework that integrated service restoration strategy by orchestrating MESS scheduling and microgrid resource dispatching while considering load consumption uncertainties. Their objective was to maximize service restoration in microgrids by effectively coordinating microgrid resource dispatching and MESS scheduling, with MESS fleets being dynamically dispatched among microgrids to facilitate load restoration in tandem with microgrid operation. To optimally schedule these processes, they employed a DRL algorithm, specifically the twin delayed deep deterministic policy gradient (TD3), which facilitated the training of deep Q-networks and policy networks. The well-trained policy network was then applied online to execute multiple actions simultaneously. Their proposed model was assessed using an integrated test system comprising three microgrids interconnected by the Sioux Falls transportation network. Simulation results underscored the successful coordination of mobile and stationary energy resources in enhancing system resilience.

In [73], the authors focused on enhancing grid resilience by utilizing DERs in distribution systems to restore critical loads following extreme events. They recognized the complexity of coordinating multiple DERs in a sequential restoration process, especially in the presence of uncertainties related to renewable energy sources and fuel availability. To address this challenge, the researchers turned to RL due to its capability to handle system nonlinearity and uncertainty effectively. RL's ability to be trained offline and provide immediate actions during online operations made it well suited for time-sensitive scenarios like load restoration. Their study centered on prioritized load restoration within a simplified distribution system, considering imperfect renewable generation forecasts, and compared the performance of an RL controller with that of a deterministic model predictive control (MPC) approach. The results demonstrated that the RL controller, by learning from experience and adapting to imperfect forecasts, offered a more reliable restoration process compared to the baseline controller. This study underscored the potential of RL-based controllers in addressing load restoration challenges in uncertain environments, particularly when coupled with high-performance computing (HPC), showcasing their effectiveness in the power system domain.

Reference [88] focused on enhancing the resilience of microgrids through the coordinated deployment of mobile power sources (MPSs) and repair crews (RCs) following extreme events. Unlike previous centralized approaches, which assumed uninterrupted communication networks, this research adopted a decentralized framework to address real-world scenarios where communication infrastructure might be compromised. They introduced a two-level hierarchical MARL method. At the high level, this approach determined when to prioritize power or transport networks, while the low level handled scheduling and routing within these networks. To improve learning stability and scalability, an embedded function capturing system dynamics was incorporated. Case studies using power networks validated the method's effectiveness in microgrid load restoration. This work represents a significant advancement in decentralized coordination for enhancing microgrid resilience, offering privacy preservation and robustness while outperforming existing centralized and decentralized methods.

3.3. Energy Management and Control

Energy management (EM) and adaptive control within the context of resilient power and energy systems are integral strategies and methodologies employed to enhance the reliability, efficiency, and robustness of energy distribution and consumption. Energy management encompasses a range of practices that involve monitoring, optimizing, and controlling various aspects of energy usage, generation, and distribution. These practices aim to achieve multiple objectives, including minimizing energy costs, reducing peak loads, maintaining a balance between electricity supply and demand, and ensuring the stable

operation of energy systems. EM plays a crucial role in enhancing resilience by allowing for proactive responses to disruptions, optimizing resource allocation, and minimizing the impact of unforeseen events [70]. Adaptive control, on the other hand, refers to the ability of an energy system to autonomously adjust its operation in real time based on changing conditions and requirements. It involves the use of feedback mechanisms, data analytics, and control algorithms to continuously monitor system performance, detect anomalies or faults, and make rapid and informed decisions to maintain system stability and reliability. Adaptive control is essential in resilient energy systems as it enables them to self-regulate, adapt to dynamic situations, and recover quickly from disturbances or failures [89]. Together, energy management and adaptive control form a dynamic framework that ensures the efficient use of energy resources, maintains grid stability, and responds effectively to various challenges, including load variations, demand fluctuations, cyber threats, equipment faults, and other disruptions. These strategies are critical for enhancing the resilience and sustainability of power and energy systems in the face of evolving complexities and uncertainties. Table 6 presents a summary of papers on DRL applications in energy management and control aspects of resilient power and energy systems.

Table 6. Summary of papers on DRL applications in energy management and control.

Paper (Authors and Reference)	DRL Algorithm	Main Contributions
Deshpande et al. [90]	PPO	Focus on energy management within microgrids, addressing renewable energy source variability.
Zhang et al. [91]	Bayesian DDPG	Proposed Bayesian DDPG for resilient multi-energy microgrid control.
Wang et al. [92]	Dueling DQN	Introduced Dueling DQN-based DRL for incentive demand response with interruptible load.
Raman et al. [93]	Q-learning	Compared RL-based controller to MPC for resilient energy supply during disasters.
Tightiz et al. [94]	DDPG and SAC	Applied DRL for microgrid energy management and participation in the electricity market.
Hasan et al. [95]	Q-learning	Developed control strategy for a resilient community microgrid.

In [90], the authors focused on energy management within microgrids, anticipating the growing role of small-scale renewable energy sources like photovoltaic panels and wind turbines. To address the inherent variability of renewables, the researchers applied MARL, with each agent representing a component of the microgrid. These agents were trained to autonomously optimize energy distribution, leveraging historical energy consumption and renewable production data. The simulation results demonstrated effective energy flow management, and a quantitative evaluation compared the approach to linear programming solutions. Furthermore, the study emphasized decentralization, envisioning systems capable of independently responding to grid disturbances. Real-world energy data and input from industrial users contributed to the model's development, and a generalization method was introduced to enhance its adaptability, ultimately leading to improved resilience and reliability in microgrid energy management compared to models trained on specific data.

In [91], the authors addressed the global shift towards cleaner energy systems with increased reliance on renewable energy sources (RESs) and recognized the vulnerability of power systems to extreme events due to reduced backup capacity and heightened uncertainty in RES generation. To confront these challenges in multi-energy microgrids, a Bayesian DRL approach was proposed. Unlike traditional deterministic RL, this approach

incorporated a Bayesian probabilistic network to approximate the value function distribution, mitigating the Q-value overestimation problem. The study compared this Bayesian approach, known as Bayesian deep deterministic policy gradient (BDDPG), with the conventional DDPG and optimization methods across various operational scenarios. Case studies revealed that BDDPG, by utilizing the Monte Carlo posterior mean of the Bayesian value function distribution, could achieve near-optimal policies with enhanced stability, underscoring its robustness and practicality in resilient multi-energy microgrid control, particularly in the face of uncertainties associated with extreme events.

In [92], the authors focused on enhancing incentive demand response (DR) with interruptible load (IL), which allows for swift response and improved demand-side resilience. Conventional model-based optimization algorithms for IL necessitate explicit system models, posing challenges in adapting to real-world operational conditions. Therefore, a model-free DRL approach, employing the dueling deep Q-network (DDQN) structure, was introduced to optimize IL-driven DR management under time-of-use (TOU) tariffs and varying electricity consumption patterns. The authors constructed an automatic demand response (ADR) architecture based on DDQN, enabling real-time DR applications. The IL's DR management problem was formulated as an MDP to maximize long-term profit, defining state, action, and reward functions. The DDQN-based DRL algorithm effectively addressed noise and instability issues observed in traditional DQN methods, achieving the dual objectives of reducing peak load demand and operational costs while maintaining voltage within safe limits.

Ref. [93] focused on addressing the escalating need for resilient energy supply in the face of rising natural disasters, which often disrupt the conventional electrical grid. The study built upon prior research, which demonstrated that intelligent control systems could reduce the size of PV-plus-battery setups while maintaining post-blackout service quality. However, the established approach, reliant on MPC, encountered challenges related to the necessity for accurate yet straightforward models and the complexities surrounding the discrete control of residential loads. To surmount these obstacles, the paper introduced an alternative method employing RL. The RL-based controller was then rigorously compared to the previously proposed MPC system and a non-intelligent baseline controller. The results indicated that the RL controller could deliver resilient performance on par with MPC but with significantly reduced computational demands. The research centered on a single-family dwelling equipped with solar PV panels, a battery storage system, and three distinct loads, with a primary focus on preserving refrigerator temperature and prolonging battery life during extended power outages. Despite the inherent challenges, including state constraints, the RL-based formulation was supposed to effectively address these issues, demonstrating its potential for robust energy management in the face of grid disruptions.

In [94], the authors focused on addressing the challenges associated with integrating renewable energy resources into microgrid energy management systems (EMSs) while participating in the electricity market and providing ancillary services to the utility grid. To tackle these complexities, the researchers deployed DDPG and SAC methods within an RL framework. This approach aimed to optimize the microgrid's energy management in a high-dimensional, continuous, and stochastic environment. Additionally, the microgrid was designed to act as a participant in the power system integrity protection scheme, responding promptly to utility grid protection requirements using its available resources. A real-world dataset was used to validate the proposed methods. The key contributions of the paper included defining the microgrid's structure, elements, and constraints for the MDP, introducing a DRL framework for microgrid EMS using DDPG and SAC, and evaluating the technique's performance in both normal and contingency scenarios.

In [95], a control strategy for a resilient community microgrid was developed. This microgrid model incorporated solar PV generation, electric vehicles (EVs), and an improved inverter control system. To enhance the microgrid's ability to operate in both grid-connected and islanded modes and improve system stability, a combination of universal droop control, virtual inertia control, and RL-based control mechanisms was employed. These mechanisms

dynamically adjusted the control parameters online to fine-tune controller influence. The model and control strategies were implemented in MATLAB/Simulink and subjected to real-time simulation to assess their feasibility and effectiveness. The experimental results demonstrated the controller's effectiveness in regulating frequency and voltage under various operating conditions and microgrid scenarios, contributing to enhanced energy reliability and resilience for communities. The study also addressed the challenges posed by multiple solar PV systems and EVs in the microgrid, providing a more accurate approach to power sharing and improved stability and power quality through dynamic control adjustments.

3.4. Communications and Cybersecurity

The rise of smart grid technologies and the integration of advanced communication systems within power and energy networks have led to cybersecurity becoming a paramount concern for these systems' operators [96]. Within this context, the security of critical elements such as data availability, data integrity, and data confidentiality is seen as crucial for ensuring cyber resiliency. These fundamental elements are strategically targeted by cyber adversaries, with the aim of compromising the integrity and reliability of data transmitted across the communication networks of the power grid. The objectives pursued by these adversaries encompass a range of disruptive actions, including tampering with grid operations, the interruption of the secure functioning of power systems, financial exploitation, and the potential infliction of physical damage to the grid infrastructure. To counteract these threats, extensive research efforts have been devoted to the development of preventive measures within the realm of communications and cybersecurity. These measures are designed to deter cyber intruders from infiltrating network devices and databases [97]. The overarching goal of these preventative measures is to enhance the security posture of power and energy systems by safeguarding their communication channels and the associated cyber assets.

In this subsection, an exploration of the pertinent literature addressing the multifaceted dimensions of communications and cybersecurity within the realm of resilient power and energy systems is undertaken. Through this examination, insights are provided into the diverse strategies contributing to the protection and resilience of critical power infrastructure in the face of evolving cyber threats. Table 7 presents a summary of papers on DRL applications in the communications and cybersecurity aspects of resilient power and energy systems.

In [98], the authors focused on addressing the efficiency challenges associated with resource allocation and user scheduling within wireless networks to support near-real-time control of community resilience microgrids. To address these challenges, they introduced a DQN-based resource allocation methodology leveraging DRL. This approach aimed to optimize resource allocation for both macrocell base stations and small-cell base stations within densely populated wireless networks. The DQN scheme outperformed traditional proportional fairness (PF) and an optimization-based algorithm called distributed iterative resource allocation (DIRA) by achieving a 66% and 33% reduction in latency, respectively. Additionally, DQN demonstrated improved throughput and fairness, making it a valuable solution for latency-critical applications like future smart grids' connected microgrids. The algorithm's distributed nature reduced signaling overhead and enhanced adaptability in the network. Overall, DQN showcased its potential for various latency-sensitive applications.

In [99], the focus was on the optimal placement of phasor measurement units (PMUs) in smart grids to ensure complete system observability while minimizing the number of PMUs required. The proposed approach, termed the attack-resilient optimal PMU placement strategy, addressed the specific order in which PMUs should be placed. Using RL-guided tree search, the study employed sequential decision making to explore effective placement orders. The strategy began by identifying vulnerable buses, aiming to protect as many buses as possible during staged PMU installation to mitigate costs. This reduced the state and action space in large-scale smart grid environments. The RL-guided tree search

method efficiently determined key buses for PMU placement, resulting in a reasonable order of PMU installation. Extensive testing on IEEE standard test systems confirmed the effectiveness of this approach, demonstrating its superiority over existing methods. The study's main contributions included the introduction of a new approach that considered placement order, the use of tree search to enhance learning efficiency, and the application of a least-effort attack model for identifying vulnerable buses, making it well suited for large-scale grid environments.

Table 7. Summary of papers on DRL applications in communications and cybersecurity.

Paper (Authors and Reference)	DRL Algorithm	Main Contributions
Elsayed et al. [98]	Q-learning	Resource allocation for latency reduction in community resilience microgrids.
Zhang et al. [99]	DQN	Attack-resilient optimal PMU placement strategy.
Wei et al. [100]	DDPG	Cyber-attack recovery with optimal reclosing time.
Zhang et al. [101]	DDPG	Fuzzy-system-based DRL for demand-side management.
Etezadifar et al. [102]	DQN	Event detection in non-intrusive load monitoring with cybersecurity enhancement.
Zhang et al. [103]	Distributed DDPG	Distributed DRL for defending microgrids against FDI attacks.
Sahu et al. [104]	-	Mixed-domain RL environment for cyber-physical resilience.
Zeng et al. [105]	SAC	Adversarial MARL for cybersecurity in demand response.
Fard et al. [106]	DQN	Analysis of the resilience of data transmission when subjected to gradient-based adversarial attacks.
Chen et al. [107]	Asynchronous advantage actor–critic (A3C)	Decentralized secondary control for BESSs.
Guo et al. [108]	Minimax Q-learning	Dynamic defense against dynamic load-altering attacks.
Huang et al. [109]	-	Review of RL's role in cyber resilience and defense mechanisms.

In [100], a comprehensive strategy for mitigating the impact of cyber-attacks on critical power infrastructures was presented. The integration of cyber-physical systems has introduced vulnerabilities that can be exploited by malicious attackers, potentially disrupting power transfer by tripping transmission lines. To effectively recover from such attacks, the paper proposed an innovative recovery strategy focused on determining the optimal reclosing time for the tripped transmission lines. This strategy leveraged the power of DRL, specifically utilizing the DDPG algorithm. The continuous action space of the problem makes DDPG a suitable choice for this task. The DDPG framework consisted of a critic network and an actor network, with the former approximating the value function and the latter generating actions. To develop and train the strategy, an environment was established to simulate the power system's state transitions during cyber-attack recovery processes. Additionally, a reward mechanism based on transient energy function was used to evaluate the performance of recovery actions. Through offline training using state, action, and reward data, the DDPG-based strategy was equipped to make optimal reclosing decisions during online cyber-attack recovery. This approach was shown to outperform

existing recovery strategies that either reclose immediately or follow fixed time-delay protocols. The paper's contributions encompassed the introduction of an adaptive cyber-attack recovery strategy, the development of a simulation environment for replicating power system dynamics under attack conditions, and the ability to continuously yield optimal or near-optimal actions for different cyber-attack scenarios, effectively reducing the risks associated with cascading outages in critical power infrastructure.

In [101], a resilient optimal defensive strategy was introduced to address the challenges posed by false data injection (FDI) in demand-side management, which can impact security, voltage stability, power flow, and economic costs in interconnected microgrids. The approach utilized a Takagi–Sugeuo–Kang (TSK) fuzzy-system-based RL method, specifically employing the DDPG algorithm to train both actor and critic networks. These networks aid in security switching control strategies and multi-index assessment. An improved alternating direction method of multipliers (ADMM) method was employed for policy gradient with online coordination, ensuring convergence and optimality. Moreover, a penalty-based boundary intersection (PBI)-based multiobjective optimization technique was utilized to simultaneously address economic cost, emissions, voltage stability, and rate of change of frequency (RoCoF) limits. Simulation results validated the effectiveness of the resilient strategy in mitigating uncertain attacks on interconnected microgrids, offering an adaptable and promising solution.

In [102], an event detection algorithm based on RL was proposed for non-intrusive load monitoring (NILM) in residential applications. The method involved a feedback system that utilized several traditional event detection algorithms to train the RL agent without direct access to consumer data. The RL agent, equipped with a dual replay memory (DRM) structure, learned the knowledge of existing event detection algorithms and combined them into a versatile model. A real-world dataset was used to validate the algorithm's performance under various scenarios, including non-ideal conditions. The results showed that the proposed algorithm outperformed traditional event detection algorithms and improved the cybersecurity of participating households by isolating the agent from consumer data. The contributions of this work included the development of an RL-based event detection algorithm that excels in both ideal and non-ideal conditions, the proposal of a cybersecurity-enhancing architecture inspired by federated learning, and the introduction of a DRM structure to significantly enhance the RL algorithm performance in NILM applications.

In the context of the increasing threat of FDI attacks on the demand side of interconnected microgrids, the study by Zhang et al. [103] introduced a resilient optimal defensive strategy using distributed DRL. To assess the impact of FDI attacks on demand response, an online evaluation method employing the recursive least-square (RLS) technique was devised to gauge the effect on supply security and voltage stability. Based on this security confidence assessment, a distributed actor network learning approach was proposed to derive optimal network weights, facilitating the generation of an optimal defensive plan that addresses both economic and security concerns within the microgrids system. This methodology not only enhanced the autonomy of each microgrid but also improved DRL efficiency. Simulation results demonstrated the effectiveness of the approach in evaluating FDI attack impacts, highlighting the potential of an enhanced distributed DRL approach for robustly defending microgrids against demand-side FDI attacks. The key contributions of this work encompassed the development of a two-stage optimal defensive strategy, the introduction of a regularized RLS method for online FDI impact evaluation, and the proposal of a distributed RL approach to ensure microgrid economic and security resilience with improved autonomy and learning efficiency compared to existing methods.

In [104], the authors delved into the realm of data-driven approaches for controlling electric grids using machine learning, with a particular focus on RL. RL techniques, renowned for their adaptability in uncertain environments like those influenced by renewable generation and cyber system variations, present a compelling alternative to traditional optimization-based solvers. However, effectively training RL agents necessitates extensive

interactions with the environment to acquire optimal policies. While RL environments for power systems and communication systems exist separately, bridging the gap between them in a unified, mixed-domain cyber-physical RL environment has been a significant challenge. Existing co-simulation methods, while efficient, demanded substantial resources and time for generating extensive datasets to train RL agents. Therefore, this study concentrated on crafting and validating such a mixed-domain RL environment, utilizing OpenDSS for power systems and SimPy, a versatile discrete event simulator in Python, for cyber systems, ensuring compatibility across different operating systems. The primary objective was to empower the distribution feeder system to exhibit resilience in both the cyber and physical domains. This was achieved through the training of agents for tasks like network reconfiguration, voltage control, and rerouting, thereby minimizing the steps required to restore communication and power networks in the face of various threats and contingencies. The key contributions of this paper included the development of a discrete event simulation-based cyber RL environment for training agents, the creation of an OpenDSS-based RL environment for distribution grid reconfiguration, the integration of SimPy and OpenDSS for comprehensive cyber-physical defense, validation across different power and cyber system sizes, and the successful training of well-known RL agents using these environments, applying the MDP model to rerouting and network reconfiguration tasks in their respective environments.

In [105], the role of demand response in enhancing grid security by maintaining the demand–supply balance in real time through consumer flexibility adjustments was explored. The proliferation of digital communication technologies and advanced metering infrastructures has led to the adoption of data-driven approaches, such as MARL, for solving demand response challenges. However, the increased data interactions within and outside the demand response management system have introduced significant cybersecurity threats. This study aimed to address the cybersecurity aspect by presenting a resilient adversarial MARL framework for demand response. The framework constructed an adversary agent responsible for formulating adversarial attacks to achieve worst-case performance. It then employed periodic alternating robust adversarial training with the optimal adversary to mitigate the impacts of adversarial attacks. Empirical assessments conducted within the CityLearn Gym environment highlighted the vulnerability of MARL-based demand response systems to adversary agents. However, the proposed approach exhibited substantial improvements in system resilience, reducing net demand ramping by approximately 38.85%. The work introduced new adversarial training methods, addressed robustness challenges, and provided insights into the impact of pre-training on control policies, contributing significantly to enhancing the cybersecurity of MARL-based demand response systems.

In [106], an investigation was conducted into the resilience of data transmission between agents in a cluster-based, heterogeneous, MADRL system when subjected to gradient-based adversarial attacks. To address this challenge, an algorithm utilizing a DQN approach, in combination with a proportional feedback controller, was introduced to enhance the defense mechanism against fast gradient sign method (FGSM) attacks and improve the performance of DQN agents. The feedback control system served as a valuable auxiliary tool for mitigating system vulnerabilities. The resilience of the developed system was evaluated under FGSM adversarial attacks, categorized into robust, semi-robust, and non-robust scenarios based on average reward and DQN loss. Data transfers were examined within the MADRL system, considering both real-time and time-delayed interactions, in leaderless and leader–follower scenarios. The contributions of this research included the design of a proportional controller to bolster the DQN algorithm against FGSM adversarial attacks, the exploration of on-time and time-delayed data transmissions to enhance the defense strategy, and the demonstration of the superior performance achieved by integrating the proportional controller into the DQN learning process, resulting in higher average cumulative team discounted rewards for the MADRL system.

In [107], a decentralized secondary control scheme was presented for multiple heterogeneous BESSs within islanded microgrids. Unlike prior approaches, that involve extensive information transmission among secondary controllers, this scheme eliminated the need for excessive real-time data exchange, reducing communication costs and minimizing vulnerability to cyber-attacks. The secondary control method simultaneously achieved frequency regulation and state-of-charge (SoC) balancing for BESSs, all without necessitating precise BESS models. This was achieved through an asynchronous advantage actor–critic (A3C)-based MADRL algorithm, featuring centralized offline learning with shared convolutional neural networks (CNNs) to maximize global rewards. A decentralized online execution mechanism was employed for each BESS. Additionally, to counter potential denial-of-service (DoS) attacks on local communication networks, a signal-to-interference-plus-noise ratio (SINR)-based dynamic and proactive event-triggered communication mechanism was introduced, reducing the impact of DoS attacks and conserving communication resources. Simulation results demonstrated that the proposed decentralized secondary controller effectively accomplishes simultaneous frequency regulation and SoC balancing. Comparative analysis against other event-triggered methods and MA-DRL algorithms highlighted the superiority of the A3C-based MA-DRL algorithm with CNN, capable of adapting its release frequency based on real-time SINR to mitigate network bandwidth occupation and packet loss rates induced by DoS attacks. The paper contributed to the development of a data-driven decentralized secondary control system for microgrids with multiple heterogeneous BESSs, offering a new approach to address these challenges.

In [108], a dynamic defense strategy was introduced to counter dynamic load-altering attacks (D-LAAs) in the context of cyber-physical threats to interconnected power grids. Unlike traditional static defense approaches, this strategy has been designed to consider a multistage game between the attacker and defender, where both parties' actions evolve dynamically. Minimax Q-learning was applied to determine the optimal strategies at each state, with the attacker adjusting their actions based on feedback, particularly the cascading failure and load shedding measurements. The proposed model's effectiveness was assessed using the IEEE 39-bus system, demonstrating its superiority over passive defense strategies. This dynamic defense approach was supposed to offer improved power system resilience by addressing evolving cyber-physical threats, making it a valuable preemptive strategy for safeguarding critical infrastructure. Key contributions included the extension of one-shot D-LAAs to sequential attacks, a two-player multistage game framework, and empirical evidence showcasing its effectiveness in reducing load losses caused by D-LAAs.

In [109], the authors addressed the growing vulnerability of cyber systems due to the increasing number of connected devices and the sophistication of cyber attackers. Traditional cybersecurity measures, such as intrusion detection and firewalls, are deemed insufficient in the face of evolving threats. Cyber resilience, as a complementary security paradigm, was introduced to adapt to both known and zero-day threats in real time, ensuring that the critical functions of cyber systems remain intact even after successful attacks. The cyber-resilient mechanism (CRM) has been central to this concept, relying on feedback architectures to sense, reason, and act against threats. RL plays a vital role in enabling CRMs to provide dynamic responses to attacks, even with limited prior knowledge. The paper reviewed RL's application in cyber resilience and discussed its effectiveness against posture-related, information-related, and human-related vulnerabilities. It also addressed vulnerabilities within RL algorithms and introduced various attack models aimed at manipulating information exchanged between agents and their environment. The authors proposed defense methods to safeguard RL-enabled systems from such attacks. While discussing future challenges and emerging applications, the paper highlighted the need for further research in defensive mechanisms for RL-enabled systems.

3.5. Resilience Planning and Metric Development

Resilience planning in the context of power and energy systems involves planning efforts to develop comprehensive strategies for fortifying electricity infrastructure to with-

stand and recover from potential extreme events in the future [110]. It primarily focuses on identifying and prioritizing investments in the electricity grid to ensure the reliable and resilient supply of power to end-use customers. These planning-based strategies may encompass initiatives such as the installation of underground cables, strategic energy storage planning, and other infrastructure enhancements aimed at reinforcing the system's ability to deliver uninterrupted electricity [78]. On the other hand, metric development within the realm of power and energy system resilience is a foundational component for quantifying and evaluating a power and energy system's ability to endure and rebound from disruptions. These metrics serve as precise and measurable indicators, offering a quantitative means to assess the performance of a power system concerning its resilience. They provide valuable insights into how the system operates under normal conditions and its ability to withstand stressors or adverse situations. Metric development plays a crucial role in systematically gauging and enhancing the resilience of power systems, allowing for informed decision making and the optimization of infrastructure investments. Table 8 presents a summary of papers on DRL applications in resilience planning and metric development aspect of resilient power and energy systems.

Table 8. Summary of papers on DRL applications in resilience planning and metric development.

Paper (Authors and Reference)	DRL Algorithm	Main Contributions
Pang et al. [111]	-	Incorporating battery degradation into microgrid expansion planning.
Pang et al. [112]	Double DQN	Cost-effective microgrid expansion considering uncertainties.
Paul et al. [113]	Q-learning	Risk-based approach for enhancing distribution grid resilience.
Ibrahim et al. [114]	DQN, Double DQN, and REINFORCE	Introducing level-of-resilience (LoR) metric and evaluating power system resiliency.

In [111], the primary objective was to enhance power resilience while minimizing overall costs in long-term microgrid expansion planning. Unlike the existing literature, this study incorporated the real-world battery degradation mechanism into the microgrid expansion planning model. The approach involved employing RL-based simulation methods to derive an optimal expansion policy for microgrids. Through case studies, the effectiveness of this model was confirmed, and the impact of battery degradation was investigated. Additionally, the paper explored how the unavailability of power plants during extreme outages affected optimal microgrid expansion planning. Battery capacity naturally diminishes over time and use due to chemical reactions within the battery, ultimately leading to its disposal. Considering this battery degradation in storage-expansion planning for microgrids, the paper introduced a long-term expansion planning framework using a DRL algorithm and simulation-based techniques.

In [112], the authors focused on developing a model for long-term microgrid expansion planning using various DRL algorithms including DQN, Double DQN, and REINFORCE. This model introduced multiple energy resources and their associated uncertainties, such as battery cycle degradation. The study employed a DRL approach to derive cost-effective microgrid expansion strategies aimed at enhancing power resilience, particularly during grid outages. Through case studies, the paper demonstrated the effectiveness of this approach, showcasing how it optimized microgrid expansion planning while accounting for factors like battery degradation and resilience constraints. The proposed method offered valuable insights into creating backup power solutions for customers during grid

disruptions, considering the real-world characteristics and uncertainties associated with various power generation and energy storage units.

In [113], the authors focused on enhancing the resilience of power distribution systems against extreme events, particularly high-impact low-probability (HILP) incidents, through the optimization of grid-hardening strategies. The study employed a risk-based metric for quantifying resilience and used a Q-learning algorithm to determine the optimal sequence of grid-hardening actions while adhering to a budget constraint. The objective was to minimize the conditional value at risk (CVaR) for the loss of load. The research demonstrated the practical application of this approach through a case study involving the IEEE 123-bus test feeder, highlighting its effectiveness in strategically allocating limited resources for resilient distribution system planning. This study contributed to the field by providing a risk-aware framework for enhancing distribution grid resilience, with a specific focus on undergrounding distribution lines to withstand wind storms, offering valuable insights for power distribution system operators aiming to strengthen their networks against extreme events.

In [114], an assessment of power system resilience was performed by introducing a metric called the level of resilience (LoR), which quantified system resilience in terms of the minimum number of faults required to induce a blackout through sequential topology attacks. The study employed four DRL methods, namely, DQN, double DQN, REINFORCE, and REINFORCE with baseline, to determine the LoR. The research conducted three case studies using the IEEE 6-bus test system, focusing on evaluating the agents' performance. Notably, the double DQN agent excelled by achieving the highest success rate and demonstrating superior efficiency compared to the other agents. This work's main contribution was to utilize DRL techniques to evaluate power system resilience by determining the minimum fault count necessary for a system blackout under sequential topological attacks, offering valuable insights for system designers in selecting the most resilient topology.

4. Challenges, Limitations, and Future Research Directions

In this section, the challenges and limitations encountered when applying DRL for resilient power and energy systems are confronted. Potential future research directions aimed at addressing these challenges and maximizing the effectiveness of DRL are envisioned in enhancing power and energy system resilience.

4.1. Challenges and Limitations

The integration of DRL into resilient power and energy systems, while promising, is not without its set of challenges and limitations. Understanding these hurdles is crucial for devising effective strategies and solutions. Here, these challenges and limitations are discussed in detail.

4.1.1. Sub-Optimal Solutions in DRL Applications

A significant and recurrent challenge within the domain of DRL applications is the propensity to produce sub-optimal solutions. This issue arises when DRL agents have not undergone sufficient training or when the delicate balance between exploration and exploitation is not adequately maintained. The consequences of sub-optimal solutions can be particularly critical in scenarios where resilient decision making is paramount. To elucidate this challenge, insights can be drawn from recent research.

The work by Abdelmoaty et al. [115] provides valuable insights into the sub-optimality concern. In their study, the authors conducted a comparative analysis between two approaches: resilient topology design for wireless backhaul using integer linear programming (ILP) and employing a DRL-based method. Their findings shed light on the sub-optimal nature of solutions derived from DRL techniques. This research underscores the need for careful consideration when applying DRL in the context of enhancing resilience in power and energy systems.

Moreover, the study conducted by Nguyen et al. [116] emphasizes that in complex and intricate environments, DRL agents are susceptible to becoming ensnared in sub-optimal solutions. This highlights the significance of exploring alternative training methodologies to ensure that DRL agents are equipped to discover optimal solutions even in intricate scenarios.

Addressing the challenge of sub-optimal solutions in DRL applications demands meticulous attention and innovative approaches. Striking a balance between exploration and exploitation, optimizing training regimes, and continuously refining DRL algorithms are pivotal steps toward mitigating this limitation. As we advance in harnessing the potential of DRL for resilience enhancement in power and energy systems, an acute awareness of the sub-optimality challenge will be instrumental in driving progress and ensuring the reliability of critical infrastructure.

4.1.2. Addressing Scalability Challenges in DRL Applications

One of the formidable challenges encountered in applying DRL to power and energy systems lies in scalability. The inherent complexities of these systems can result in exceedingly large state spaces, rendering the training process computationally demanding and time consuming. Tackling scalability concerns becomes imperative to harness the full potential of DRL in this domain. A deeper examination of this challenge, along with potential solutions, can provide valuable insights.

An illustrative example of scalability concerns is discussed in the work by Sami et al. [117], where they specifically address the issue of action space size and the limitations associated with tabular Q-learning in the context of MDP design. This research not only highlights the challenges posed by scalability but also demonstrates that careful design considerations can mitigate these concerns. By presenting a practical implementation example related to fog and service placement problems, they showcase how thoughtful DRL solutions can effectively address scalability hurdles.

In the realm of multi-agent DRL, scalability challenges take on a different dimension. Qu et al. [118] delve into this aspect by emphasizing that even when individual agents have relatively small state or action spaces, the global state or action space can grow exponentially with the number of agents. This exponential explosion in complexity necessitates innovative techniques to enhance scalability in multi-agent DRL scenarios.

Applying these insights to the domain of resilient power and energy systems, it becomes evident that a nuanced and strategic approach is vital in addressing scalability concerns. Tailoring DRL algorithms to accommodate the intricacies of these systems, optimizing the representation of state and action spaces, and exploring innovative techniques for multi-agent scenarios are all critical steps. In this way, researchers and practitioners can unlock the potential of DRL to efficiently navigate the vast and complex landscapes of power and energy infrastructure, ultimately enhancing resilience and reliability.

4.1.3. Navigating the Time–Accuracy Balance in DRL Applications

In the realm of DRL, a compelling aspect is its capacity to make swift decisions, often outpacing traditional optimization methods. However, this agility comes with a noteworthy tradeoff between speed and accuracy. While DRL can excel in delivering rapid responses, achieving the utmost level of precision may necessitate extended training periods and exploration phases. It is crucial to understand this tradeoff and its implications, especially when applying DRL to enhance the resilience of power and energy systems.

In situations where immediate decisions are paramount, DRL can prove advantageous. For instance, in the context of cloud robotics, as demonstrated by Penmetcha et al. [119], DRL-based dynamic computational offloading methods can yield rapid decisions, with a mean computation time of 71.28 milliseconds, while achieving a commendable final accuracy of 84%. This showcases the potential of DRL in scenarios where timely responses are essential.

However, it is equally important to acknowledge that certain applications within the realm of power and energy systems may prioritize accuracy over speed. In critical situations, where the resilience and reliability of the energy infrastructure are at stake, precision becomes paramount. In such cases, sacrificing accuracy for expediency may not be a viable option, and DRL may not align with the requirements of the application.

Thus, striking the right balance between time and accuracy is a pivotal consideration when employing DRL in the context of power and energy systems' resilience. Careful evaluation of the specific demands of each application, along with a nuanced understanding of the tradeoffs involved, will guide the judicious use of DRL, ensuring that it aligns with the objectives and constraints of the given scenario.

4.1.4. Navigating Complexity with Model-Free DRL Algorithms

In the landscape of RL, several dimensions of complexity have been explored, encompassing space complexity, computational complexity, and sample complexity, as elucidated in the comprehensive work by Strehl et al. [120]. Within this intricate landscape, a specific category known as model-free RL algorithms emerges, offering distinctive advantages and challenges.

Strehl et al. [120] provide valuable insights into what constitutes a model-free RL algorithm. Specifically, a model-free RL algorithm is characterized by its space complexity, which is asymptotically lower than the space required to store a Markov decision process (MDP). This succinct definition encapsulates the essence of model-free RL, emphasizing its fundamental departure from traditional RL paradigms that rely on MDPs for decision making.

The model-free approach of DRL assumes paramount significance when dealing with the intricacies of power and energy systems. These systems often defy precise modeling due to their inherent complexities, dynamic nature, and the influence of external factors. In such scenarios, attempting to construct an accurate MDP becomes a formidable challenge, if not an impossibility. This is precisely where model-free DRL shines, as it does not depend on a predefined MDP structure.

However, the adoption of a model-free stance introduces its own set of complexities. For the DRL agent to navigate and make decisions effectively, it must essentially start from scratch. This implies that the agent undertakes a process of learning by trial and error, relying on interactions with the environment to glean insights and refine its decision making. This iterative learning process can be computationally demanding and often necessitates copious amounts of training data.

The computational complexity inherent in model-free DRL can pose challenges, especially in resource-constrained environments or real-time applications. Moreover, the agent's ability to generalize from its experiences and effectively explore the vast state-action space can be a delicate balancing act, requiring careful consideration.

In a nutshell, model-free DRL adds a new level of complexity while liberating decision making from the restrictions of explicit MDP modeling. Model-free DRL's benefits and computational requirements must be balanced in order for it to be successfully applied in the field of resilient power and energy systems. To fully utilize the promise of model-free DRL in improving the robustness and effectiveness of these crucial systems, careful consideration of data needs, training methodologies, and computational resources is important.

4.1.5. Safety Considerations

The integration of DRL algorithms in the context of enhancing the resilience of power and energy systems necessitates robust safety measures. The implementation of RL techniques introduces various safety concerns that warrant careful consideration. The authors in [121] conducted a comprehensive examination of safety-related issues associated with RL. In their study, they delve into the realm of safety policies, safety complexity, safety applications, safety benchmarks, and safety challenges concerning RL.

One of the primary concerns when employing DRL in power and energy systems is the potential for incorrect actions or policies generated during training or deployment. These erroneous decisions have the potential to result in system instability or even physical damage. Therefore, ensuring the safety and reliability of DRL-driven decisions within the context of resilient power and energy systems emerges as a critical and non-negotiable challenge. Safeguarding against adverse outcomes and unforeseen consequences remains a paramount consideration in the deployment of DRL applications, especially in scenarios where the consequences of failure can be severe and far-reaching. Consequently, addressing safety concerns is an imperative aspect of successfully implementing DRL applications in this domain.

4.1.6. Generalization Challenges

Despite the impressive capabilities exhibited by state-of-the-art DRL algorithms in solving complex tasks, their ability to generalize between tasks and adapt to new environments remains a formidable challenge. This limitation becomes apparent when considering the difficulty of transferring an agent's learned experience to novel situations, as discussed by [122]. While RL agents may excel in mastering multiple levels of a video game, they often face catastrophic failures when confronted with a slightly different, previously unseen level. In stark contrast, humans exhibit the remarkable ability to seamlessly generalize their knowledge and skills across similar tasks, a capacity that remains largely absent in RL agents. This phenomenon results in RL agents becoming overly specialized in the specific environments encountered during their training, making them less adaptable to new challenges and hindering their ability to generalize effectively [122].

Furthermore, the study conducted by [123] underscores the challenges faced by deep RL in terms of generalization beyond the scope of their training environments. To address this issue, the researchers introduce a benchmark and experimental protocol designed to assess the generalization capabilities of deep RL algorithms. Their extensive evaluation reveals that, in some cases, standard deep RL algorithms outperform specialized ones explicitly designed for generalization. This suggests that while RL has achieved significant success in various tasks, its ability to generalize effectively to diverse and previously unseen scenarios remains a research frontier. In the context of resilient power and energy systems, where adapting to a wide range of unforeseen challenges is crucial, the capacity of DRL models to generalize and provide effective solutions is of paramount importance and an ongoing area of investigation.

4.1.7. Ethical Considerations

The integration of DRL into resilient power and energy systems raises significant ethical considerations that must be addressed. DRL systems, often operating as complex neural networks, pose challenges related to transparency and accountability [124]. Ensuring that these systems' decision-making processes are transparent and that there is accountability for their actions is crucial to identify and rectify issues effectively. Additionally, fairness and equity are paramount concerns, as unintended biases in training data or algorithms can result in unfair treatment and unequal resource distribution [125].

Protecting privacy and data security is both a legal and ethical obligation, particularly when DRL systems rely on sensitive information [126]. Furthermore, ethical DRL applications should prioritize environmental sustainability, incorporating eco-friendly practices, and allow human experts to intervene or override automated decisions in critical situations. Resource allocation guided by ethical principles ensures fair and efficient distribution of resources. Addressing these ethical concerns is essential to build trust and ensure the responsible and ethical deployment of DRL in resilient power and energy systems.

4.2. Future Research Directions

Exploring the potential of DRL in enhancing the resilience of power and energy systems has opened up a myriad of opportunities and challenges. As DRL continues to

evolve and mature as a field, there are several promising avenues for future research that can further advance the application of DRL in resilient power and energy systems. These directions not only aim to address existing limitations but also seek to leverage the unique capabilities of DRL to transform the way we manage and optimize energy infrastructure in the face of disruptions and uncertainties. In this subsection, some key areas are outlined where future research in this domain can make significant contributions.

4.2.1. Development of Resilience Metrics

Quantifying and assessing the resilience of power and energy systems is fundamental for effective decision making during extreme events or disruptions. As these systems are increasingly exposed to a wide range of challenges, the development of robust resilience metrics becomes paramount. DRL offers a promising avenue for the advancement of such metrics. Building upon the foundations laid by researchers like those in [114], who have introduced novel approaches for measuring system resilience in different contexts, there is an opportunity to extend this work to the realm of power and energy systems.

Future research can focus on harnessing the capabilities of DRL to create sophisticated resilience metrics tailored to the specific challenges faced by energy infrastructure. These metrics can encompass a variety of factors, including system topology, fault tolerance, response times, and resource allocation. By leveraging DRL algorithms, which excel at learning and adapting in dynamic environments, it becomes possible to develop metrics that evolve and adapt as the system's conditions change.

Moreover, the integration of real-time data streams into DRL-based resilience metrics can enable continuous monitoring and assessment, providing operators with timely insights into the system's resilience status. This fusion of machine learning and resilience assessment has the potential to revolutionize how we gauge the robustness of power and energy systems, ultimately leading to more informed decision making and enhanced preparedness for future challenges. Therefore, the development of resilience metrics through DRL stands as a crucial and promising avenue for future research in this field.

4.2.2. Enhancing Training Strategies

Enhancing the training process of DRL approaches holds significant promise for improving their applicability in resilient power and energy systems. Researchers, as demonstrated in [127], have explored asynchronous variants of conventional RL algorithms. These variants leverage parallel actor–learners to stabilize the training process, thereby enhancing the effectiveness of various RL methods in training neural networks. By investigating similar techniques in the context of power and energy systems, we can potentially mitigate some of the challenges posed by large-scale and complex environments.

Curriculum learning, as introduced in [128], presents another avenue for refining DRL training. This methodology focuses on optimizing the order in which the agent accumulates experiences, aiming to boost overall performance and training speed on a predefined set of tasks. In the context of power and energy systems, designing curricula that expose DRL agents to progressively more complex scenarios could expedite learning and result in more efficient and resilient systems.

Additionally, transfer learning, also known as knowledge transfer, is emerging as a critical technique in RL, as highlighted in [129]. This approach leverages external expertise to enhance the learning process in a target domain. In the context of resilient energy systems, the application of transfer learning could involve adapting pre-trained DRL models from related domains to expedite the training and deployment of agents in energy-related tasks.

Furthermore, meta-learning, as discussed in [130], involves optimizing meta-parameters in RL algorithms to facilitate more efficient learning. Future research should delve into these innovative training techniques, tailoring them to the unique challenges of power and energy systems. In this way, we can potentially accelerate the convergence of DRL algorithms, making them more adaptable and effective in managing the resilience of large-

scale energy infrastructure. This research direction holds promise for enhancing the speed and efficiency of DRL-based solutions, ensuring their practicality in real-world scenarios.

4.2.3. Enhancing Transparency with Explainable AI (XAI) in DRL

In the realm of DRL, the adoption of explainable AI (XAI) principles emerges as a crucial endeavor. XAI aims to address the pressing issues of trust and confidence in AI systems, especially in scenarios where safety considerations are paramount [131]. This aspect assumes heightened significance when applied to power and energy systems, where operational reliability and resilience are non-negotiable.

A thought-provoking article by [132] underscores the pivotal role of XAI technology within the domain of DRL models for power system emergency control. The core motivation here is to establish transparency and trust in AI-driven decision-making processes, a prerequisite for effective resilience enhancement. The article introduces a method known as Deep-SHAP, designed explicitly to inject interpretability into these complex DRL models.

By integrating XAI principles into DRL models, the way can be paved for an enhanced understanding of AI-driven decisions. This newfound transparency empowers power system operators and stakeholders, enabling them to grasp the rationale behind AI-generated actions. Consequently, the decision-making process becomes more understandable and trustworthy, fostering collaborative efforts to bolster the resilience of power and energy systems. As we delve deeper into this research direction, we hold the potential to bridge the gap between advanced AI techniques and human oversight, ultimately advancing the robustness and dependability of energy infrastructure in the face of adversity.

4.2.4. Leveraging Hybrid Approaches for Enhanced Resilience

In the pursuit of fortifying the resilience of power and energy systems, it becomes increasingly evident that a one-size-fits-all solution may not suffice. Instead, adopting hybrid approaches that seamlessly blend various optimization techniques with DRL emerges as a promising avenue.

A notable illustration of this concept can be found in the work of [133], where they introduce a novel hybrid approach. Their methodology combines the prowess of stochastic programming with the adaptability of DRL to tackle the intricate challenge of Volt-VAR optimization within active distribution systems. By synergizing these diverse techniques, they enhance the system's ability to maintain voltage levels while minimizing power losses—a testament to the potential of hybridization.

Future research endeavors should delve into the uncharted territory of hybrid approaches, seeking to unlock the synergistic potential lying at the intersection of DRL and traditional optimization methods. In this way, we can harness the strengths of both worlds, augmenting the robustness and efficiency of resilience strategies for power and energy systems. These hybrid methodologies promise to offer tailored solutions that can adeptly navigate the multifaceted landscape of resilience challenges, ultimately ensuring the dependable operation of critical infrastructure during adverse conditions.

4.2.5. Human-in-the-Loop DRL: A Fusion of Human Expertise and Machine Learning

Human-in-the-loop DRL has been gaining traction across various domains, showcasing its potential to revolutionize learning paradigms. A pioneering work by [134] introduced the concept of "protocol programs," an agent-agnostic framework tailored for human-in-the-loop reinforcement learning. This innovative schema seeks to augment learning by seamlessly integrating human guidance, all without imposing rigid constraints on agent representations. In a compelling demonstration, the authors illustrated that established techniques like action pruning, reward shaping, and simulation-driven training can be elegantly encapsulated as special instances of this versatile framework. Their preliminary experiments in relatively straightforward domains illuminate the promise of this approach.

Moreover, the realm of continuous action spaces witnessed a breakthrough in the paper by [135], which introduced the Q-value-dependent policy-based human-in-the-loop reinforcement learning (QDP-HRL) algorithm. This novel approach, seamlessly integrated with the twin delayed deep deterministic policy gradient algorithm (TD3), employed human expertise judiciously. During the initial stages of learning, the human expert selectively provided guidance, guided by discrepancies in the twin Q-networks' output. To further bolster learning efficacy and performance in diverse continuous action space tasks, the authors devised an advantage loss function, drawing insights from both expert experience and agent policies.

As the field of DRL continues to evolve, one promising avenue for future research lies in the seamless integration of human expertise into resilience frameworks for power and energy systems. This collaborative approach aims to forge synergistic human–machine partnerships, capitalizing on the unique strengths of each component. In this way, we can envision a future where human intuition and domain knowledge harmoniously merge with the adaptive capabilities of DRL, paving the way for more resilient, efficient, and trustworthy energy infrastructure.

4.2.6. Setting Standards in Benchmarking and Evaluation

In the realm of RL, research by [136] has unveiled continual reinforcement learning agents (CORA), a pioneering platform designed to propel the field forward. CORA's significance lies in its ability to surmount obstacles in continual RL, offering a holistic solution encompassing benchmarks, metrics, and baselines, all conveniently encapsulated within a unified code package. This transformative platform facilitates rigorous evaluations spanning various dimensions of continual RL, operating seamlessly across diverse environments and tasks. In essence, CORA acts as a catalyst, nurturing the growth in novel algorithms within the continual RL community.

Drawing inspiration from this paradigm-shifting approach, a promising future research avenue emerges for the field of DRL applied to power and energy system resilience. Herein, the proposal is to establish standardized benchmarks and evaluation metrics tailor-made for DRL applications within this critical domain. In this way, the research community can pave the way for equitable comparisons and foster advancements in the field.

Much like CORA, these benchmarks and metrics would serve as essential tools, enabling researchers and practitioners to gauge the effectiveness of DRL algorithms within power and energy systems. By providing a common ground for assessment, such standardized measures would not only enhance the transparency and reproducibility of research but also facilitate the identification of cutting-edge solutions and best practices. Ultimately, this endeavor holds the potential to accelerate the development of robust, efficient, and resilient energy infrastructure, benefiting society as a whole.

4.2.7. Physics-Inspired Reinforcement Learning

A study by [137] introduces an innovative fusion of physics-inspired principles and DRL to tackle the intricate challenges of optimizing DERs within modern power systems. The research showcases a new architecture encompassing a graph convolutional neural network (GCN) combined with RL techniques. This dynamic approach focuses on training online controller policies, particularly targeting Volt/Var and Volt/Watt control logic within smart inverters.

Intriguingly, the study not only highlights the efficacy of the GCN-based framework in voltage regulation but also underscores its capability to mitigate voltage dynamics induced by cyber-attacks. Moreover, its robustness in the face of dynamic changes in grid configurations and its potential for transfer learning make it a powerful tool for enhancing power system resilience.

Building upon this innovative approach, the application of physics-inspired RL holds significant promise for fortifying the resilience of power and energy systems. As elucidated in Section 4.1.4, the utilization of model-free DRL can be computationally intensive and time

consuming, especially when dealing with intricate system models. Consequently, in scenarios where rapid adaptability and accurate decision making are imperative, the incorporation of model-based and physics-inspired DRL models emerges as a pragmatic solution. By seamlessly integrating domain knowledge and physical principles, such hybrid approaches can bridge the gap between computational efficiency and robust resilience, ensuring the stability and adaptability of power and energy systems in the face of diverse challenges.

4.2.8. Safe Reinforcement Learning

Safe reinforcement learning (safe RL) represents a pivotal paradigm for fortifying the resilience of power and energy systems. This approach encompasses the acquisition of policies that not only maximize return but also ensure the maintenance of reasonable system performance and adherence to vital safety constraints throughout the learning and deployment phases.

A comprehensive exploration of safe RL in [138] categorizes this domain into two distinct approaches, shedding light on their potential applications in enhancing power and energy system resilience. The first approach introduces a modification to the traditional optimality criterion, typically governed by discounted finite or infinite horizons, by incorporating a safety factor. This alteration equips RL algorithms with the capacity to consider safety as a paramount objective, thus mitigating the risk of hazardous actions.

In contrast, the second approach harnesses external knowledge or risk metrics to guide the exploration process effectively. By integrating domain-specific insights and risk-awareness metrics, this approach empowers RL agents to navigate complex environments with a heightened awareness of potential dangers and system vulnerabilities.

Considering the safety concerns discussed in Section 4.1.5, the incorporation of safe RL into the toolkit for resilient power and energy systems emerges as a compelling future research direction. By ensuring that RL-driven decisions prioritize safety alongside performance, this approach can contribute significantly to the development of robust and secure energy infrastructures capable of withstanding diverse challenges and uncertainties.

5. Summary

In this review article, a comprehensive exploration of DRL in the context of resilient power and energy systems was provided, highlighting its applications, current progress, challenges, and potential for future research. The article began by outlining the origin of DRL as a fusion of RL and deep learning, setting the foundation for understanding its multifaceted applications.

The article subsequently delved into various DRL methods and algorithms, meticulously dissecting their merits and drawbacks. This analytical foundation served as a starting point for highlighting applications of DRL across different aspects of resilient power and energy systems. These aspects included dynamic response, restoration and recovery, energy management and control, communications and cybersecurity, and resilience planning and metrics development.

One distinguishing feature of this review was its in-depth analysis of the limitations and challenges inherent in DRL. These challenges included concerns related to sub-optimality, scalability, the delicate balance between speed and accuracy, intricacies associated with system complexity, safety considerations, and the pursuit of robust generalization. This comprehensive examination not only shed light on the present hurdles but also unveiled a spectrum of future research opportunities.

The future avenues discussed encompassed enhancing training methodologies, fostering transparency through Explainable AI, leveraging hybrid approaches, integrating human expertise into DRL through human-in-the-loop paradigms, drawing inspiration from physics-based RL, and establishing the foundations of safe RL. This article equipped researchers and practitioners with a roadmap for navigating the evolving landscape of DRL in resilient power and energy systems, thereby contributing to the continual advancement of this critical field.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

A2C	Advantage actor–critic
A3C	Asynchronous advantage actor–critic
ADMM	Alternating direction method of multiplier
CNN	Convolutional neural network
DDPG	Deep deterministic policy gradient
DER	Distributed energy resource
DIRE	Disturbance and Impact Resilience Evaluation
DQN	Deep Q-network
DRQN	Deep recurrent Q-network
DRL	Deep reinforcement learning
FDI	False data injection
HILP	High-impact low-probability
IEEE	Institute of Electrical and Electronics Engineers
MADRL	Multi-agent deep reinforcement learning
MARL	Multi-agent reinforcement learning
MDP	Markov decision process
MER	Movable energy resource
MESS	Mobile energy storage system
MPC	Model predictive control
PPO	Proximal policy optimization
PMU	Phasor measurement unit
PSS/E	Power System Simulator for Engineering
RL	Reinforcement learning
SAC	Soft actor–critic
TD3	Twin delayed deep deterministic policy gradient
TRPO	Trust region policy optimization
VPG	Vanilla policy gradient

References

1. Benidris, M.; Bhusal, N.; Abdelmalak, M.; Gautam, M.; Egan, M.; Groneman, S.; Farkas, T. Quantifying Resilience Value of Solar plus Storage in City of Reno. In Proceedings of the 2021 Resilience Week (RWS), Salt Lake City, UT, USA, 18–21 October 2021; pp. 1–6.
2. Jufri, F.H.; Widiputra, V.; Jung, J. State-of-the-art review on power grid resilience to extreme weather events: Definitions, frameworks, quantitative assessment methodologies, and enhancement strategies. *Appl. Energy* **2019**, *239*, 1049–1065. [CrossRef]
3. Furman, J. *Economic Benefits of Increasing Grid Resilience to Weather Outages*; Technical Report; US Department of Energy: Washington DC, USA, 2013.
4. Smith, A.B. *U.S. Billion-Dollar Weather and Climate Disasters, 1980—Present (NCEI Accession 0209268)*; NOAA National Centers for Environmental Information: Washington, DC, USA, 2013. [CrossRef]
5. Bhusal, N.; Abdelmalak, M.; Kamruzzaman, M.; Benidris, M. Power system resilience: Current practices, challenges, and future directions. *IEEE Access* **2020**, *8*, 18064–18086. [CrossRef]
6. Mohamed, M.A.; Chen, T.; Su, W.; Jin, T. Proactive resilience of power systems against natural disasters: A literature review. *IEEE Access* **2019**, *7*, 163778–163795. [CrossRef]
7. Bhusal, N.; Gautam, M.; Abdelmalak, M.; Benidris, M. Modeling of natural disasters and extreme events for power system resilience enhancement and evaluation methods. In Proceedings of the 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Liege, Belgium, 18–21 August 2020; pp. 1–6.
8. Gautam, M.; Ben-Idris, M. Optimal Sizing of Movable Energy Resources for Enhanced Resilience in Distribution Systems: A Techno-Economic Analysis. *Electronics* **2023**, *12*, 4256. [CrossRef]

9. Nazemi, M.; Dehghanian, P.; Alhazmi, M.; Darestani, Y. Resilient operation of electric power distribution grids under progressive wildfires. *IEEE Trans. Ind. Appl.* **2022**, *58*, 1632–1643. [CrossRef]
10. Mehrjerdi, H.; Hemmati, R.; Mahdavi, S.; Shafie-Khah, M.; Catalão, J.P. Multicarrier Microgrid Operation Model Using Stochastic Mixed Integer Linear Programming. *IEEE Trans. Ind. Inform.* **2021**, *18*, 4674–4687. [CrossRef]
11. Gautam, M.; Benidris, M. Pre-positioning of movable energy resources for distribution system resilience enhancement. In Proceedings of the 2022 International Conference on Smart Energy Systems and Technologies (SEST), Eindhoven, The Netherlands, 5–7 September 2022; pp. 1–6.
12. Gautam, M.; Benidris, M. A graph theory and coalitional game theory-based pre-positioning of movable energy resources for enhanced distribution system resilience. *Sustain. Energy Grids Netw.* **2023**, *35* , 101095. [CrossRef]
13. Xie, H.; Teng, X.; Xu, Y.; Wang, Y. Optimal energy storage sizing for networked microgrids considering reliability and resilience. *IEEE Access* **2019**, *7*, 86336–86348. [CrossRef]
14. Ildarabadi, R.; Lotfi, H.; Hajiabadi, M.E. Resilience enhancement of distribution grids based on the construction of Tie-lines using a novel genetic algorithm. *Energy Syst.* **2023**, 1–31. [CrossRef]
15. Patrizi, N.; LaTouf, S.K.; Tsiropoulou, E.E.; Papavassiliou, S. Prosumer-centric self-sustained smart grid systems. *IEEE Syst. J.* **2022**, *16*, 6042–6053. [CrossRef]
16. Amarasinghe, P.G.M.; Abeygunawardane, S.K.; Singh, C. Adequacy evaluation of composite power systems using an evolutionary swarm algorithm. *IEEE Access* **2022**, *10*, 19732–19741. [CrossRef]
17. Gautam, M.; Bhusal, N.; Benidris, M. Deep Q-Learning-based distribution network reconfiguration for reliability improvement. In Proceedings of the 2022 IEEE/PES Transmission and Distribution Conference and Exposition (T&D), New Orleans, LA, USA, 25–28 April 2022; pp. 1–5.
18. Gautam, M.; Benidris, M. Distribution network reconfiguration using deep reinforcement learning. In Proceedings of the 2022 17th International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Manchester, UK, 12–15 June 2022; pp. 1–6.
19. Guo, C.; Wang, X.; Zheng, Y.; Zhang, F. Real-time optimal energy management of microgrid with uncertainties based on deep reinforcement learning. *Energy* **2022**, *238*, 121873. [CrossRef]
20. Ji, Y.; Wang, J.; Xu, J.; Fang, X.; Zhang, H. Real-time energy management of a microgrid using deep reinforcement learning. *Energies* **2019**, *12*, 2291. [CrossRef]
21. Hossain, R.; Gautam, M.; Thapa, J.; Livani, H.; Benidris, M. Deep reinforcement learning assisted co-optimization of Volt-VAR grid service in distribution networks. *Sustain. Energy Grids Netw.* **2023**, *35*, 101086. [CrossRef]
22. Hossain, R.; Gautam, M.; Lakouraj, M.M.; Livani, H.; Benidris, M. Volt-VAR optimization in distribution networks using twin delayed deep reinforcement learning. In Proceedings of the 2022 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), New Orleans, LA, USA, 24–28 April 2022; pp. 1–5.
23. Hossain, R.; Gautam, M.; MansourLakouraj, M.; Livani, H.; Benidris, M.; Baghzouz, Y. Soft Actor Critic Based Volt-VAR Co-optimization in Active Distribution Grids. In Proceedings of the 2022 IEEE Power & Energy Society General Meeting (PESGM), Denver, CO, USA, 17–21 July 2022; pp. 1–5.
24. Chen, M.; Yi, M.; Huang, M.; Huang, G.; Ren, Y.; Liu, A. A novel deep policy gradient action quantization for trusted collaborative computation in intelligent vehicle networks. *Expert Syst. Appl.* **2023**, *221*, 119743. [CrossRef]
25. Li, Y.; Wang, R.; Li, Y.; Zhang, M.; Long, C. Wind power forecasting considering data privacy protection: A federated deep reinforcement learning approach. *Appl. Energy* **2023**, *329*, 120291. [CrossRef]
26. Mahzarnia, M.; Moghaddam, M.P.; Baboli, P.T.; Siano, P. A review of the measures to enhance power systems resilience. *IEEE Syst. J.* **2020**, *14*, 4059–4070. [CrossRef]
27. Elsisi, M.; Amer, M.; Su, C.L.; Dababat, A. A comprehensive review of machine learning and IoT solutions for demand side energy management, conservation, and resilient operation. *Energy* **2023**, *281*, 128256. [CrossRef]
28. Xu, L.; Guo, Q.; Sheng, Y.; Muyeen, S.; Sun, H. On the resilience of modern power systems: A comprehensive review from the cyber-physical perspective. *Renew. Sustain. Energy Rev.* **2021**, *152*, 111642. [CrossRef]
29. Cao, D.; Hu, W.; Zhao, J.; Zhang, G.; Zhang, B.; Liu, Z.; Chen, Z.; Blaabjerg, F. Reinforcement learning and its applications in modern power and energy systems: A review. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 1029–1042. [CrossRef]
30. Perera, A.; Kamalaruban, P. Applications of reinforcement learning in energy systems. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110618. [CrossRef]
31. Zhang, Z.; Zhang, D.; Qiu, R.C. Deep reinforcement learning for power system applications: An overview. *CSEE J. Power Energy Syst.* **2019**, *6*, 213–225.
32. Xiang, X.; Foo, S. Recent advances in deep reinforcement learning applications for solving partially observable markov decision processes (pomdp) problems: Part 1—Fundamentals and applications in games, robotics and natural language processing. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 554–581. [CrossRef]
33. Vamvakas, D.; Michailidis, P.; Korkas, C.; Kosmatopoulos, E. Review and Evaluation of Reinforcement Learning Frameworks on Smart Grid Applications. *Energies* **2023**, *16*, 5326. [CrossRef]
34. Sutton, R.S.; Barto, A.G. *Introduction to Reinforcement Learning*; MIT Press: Cambridge, UK, 1998; Volume 135.
35. Zai, A.; Brown, B. *Deep Reinforcement Learning in Action*; Manning Publications: Shelter Island, NY, USA, 2020.
36. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]

37. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef]
38. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]
39. Wang, X.; Wang, S.; Liang, X.; Zhao, D.; Huang, J.; Xu, X.; Dai, B.; Miao, Q. Deep reinforcement learning: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* 2022, *early access*. [CrossRef]
40. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
41. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386. [CrossRef]
42. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009; Volume 2.
43. Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366. [CrossRef]
44. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
45. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; others. Mastering the game of Go with deep neural networks and tree search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
46. Campbell, M.; Hoane, A.J., Jr.; Hsu, F.H. Deep blue. *Artif. Intell.* **2002**, *134*, 57–83. [CrossRef]
47. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
48. Han, D.; Mulyana, B.; Stankovic, V.; Cheng, S. A Survey on Deep Reinforcement Learning Algorithms for Robotic Manipulation. *Sensors* **2023**, *23*, 3762. [CrossRef] [PubMed]
49. Sutton, R.S.; Singh, S.; McAllester, D. Comparing policy-gradient algorithms. *IEEE Trans. Syst. Man Cybern.* 2000, *early access*.
50. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.; Moritz, P. Trust region policy optimization. In Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015; pp. 1889–1897.
51. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
52. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
53. Konda, V.; Tsitsiklis, J. Actor-critic algorithms. *Adv. Neural Inf. Process. Syst.* **1999**, *12*, 1008–1014.
54. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
55. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.
56. Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. Soft actor-critic algorithms and applications. *arXiv* **2018**, arXiv:1812.05905.
57. Wong, C.C.; Chien, S.Y.; Feng, H.M.; Aoyama, H. Motion planning for dual-arm robot based on soft actor-critic. *IEEE Access* **2021**, *9*, 26871–26885. [CrossRef]
58. Hossain, R.; Gautam, M.; MansourLakouraj, M.; Livani, H.; Benidris, M. Multi-Agent Deep Reinforcement Learning-based Volt-VAR Control in Active Distribution Grids. In Proceedings of the 2023 IEEE Power & Energy Society General Meeting (PESGM), Orlando, FL, USA, 16–20 July 2023; pp. 1–5.
59. Arghandeh, R.; Von Meier, A.; Mehrmanesh, L.; Mili, L. On the definition of cyber-physical resilience in power systems. *Renew. Sustain. Energy Rev.* **2016**, *58*, 1060–1069. [CrossRef]
60. McJunkin, T.R.; Rieger, C. Resilient Control System Metrics. *Ind. Control. Syst. Secur. Resiliency Pract. Theory* **2019**, *75*, 255–276.
61. Huang, G.; Wang, J.; Chen, C.; Qi, J.; Guo, C. Integration of preventive and emergency responses for power grid resilience enhancement. *IEEE Trans. Power Syst.* **2017**, *32*, 4451–4463. [CrossRef]
62. Zhao, J.; Li, F.; Mukherjee, S.; Sticht, C. Deep reinforcement learning-based model-free on-line dynamic multi-microgrid formation to enhance resilience. *IEEE Trans. Smart Grid* **2022**, *13*, 2557–2567. [CrossRef]
63. Zhou, Z.C.; Wu, Z.; Jin, T. Deep reinforcement learning framework for resilience enhancement of distribution systems under extreme weather events. *Int. J. Electr. Power Energy Syst.* **2021**, *128*, 106676. [CrossRef]
64. Kamruzzaman, M.; Duan, J.; Shi, D.; Benidris, M. A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources. *IEEE Trans. Power Syst.* **2021**, *36*, 5525–5536. [CrossRef]
65. Chen, S.; Duan, J.; Bai, Y.; Zhang, J.; Shi, D.; Wang, Z.; Dong, X.; Sun, Y. Active power correction strategies based on deep reinforcement learning—Part II: A distributed solution for adaptability. *CSEE J. Power Energy Syst.* **2021**, *8*, 1134–1144.
66. Abdelmalak, M.; Gautam, M.; Morash, S.; Snyder, A.F.; Hotchkiss, E.; Benidris, M. Network reconfiguration for enhanced operational resilience using reinforcement learning. In Proceedings of the 2022 International Conference on Smart Energy Systems and Technologies (SEST), Eindhoven, The Netherlands, 5–7 September 2022; pp. 1–6.
67. Kadir, S.U.; Majumder, S.; Srivastava, A.; Chhokra, A.; Neema, H.; Dubey, A.; Laszka, A. Reinforcement Learning based Proactive Control for Enabling Power Grid Resilience to Wildfire. *IEEE Trans. Ind. Inform.* 2023, *early access*. [CrossRef]

68. Badakhshan, S.; Jacob, R.A.; Li, B.; Zhang, J. Reinforcement Learning for Intentional Islanding in Resilient Power Transmission Systems. In Proceedings of the 2023 IEEE Texas Power and Energy Conference (TPEC), College Station, TX, USA, 13–14 February 2023; pp. 1–6.
69. Huang, Y.; Li, G.; Chen, C.; Bian, Y.; Qian, T.; Bie, Z. Resilient distribution networks by microgrid formation using deep reinforcement learning. *IEEE Trans. Smart Grid* **2022**, *13*, 4918–4930. [CrossRef]
70. Liang, Z.; Huang, C.; Su, W.; Duan, N.; Donde, V.; Wang, B.; Zhao, X. Safe reinforcement learning-based resilient proactive scheduling for a commercial building considering correlated demand response. *IEEE Open Access J. Power Energy* **2021**, *8*, 85–96. [CrossRef]
71. Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; Zaremba, W. Openai gym. *arXiv* **2016**, arXiv:1606.01540.
72. Rieger, C.; Kolias, C.; Ulrich, J.; McJunkin, T.R. A cyber resilient design for control systems. In Proceedings of the 2020 Resilience Week (RWS), Salt Lake City, UT, USA, 19–23 October 2020; pp. 18–25.
73. Zhang, X.; Eseye, A.T.; Knueven, B.; Jones, W. Restoring distribution system under renewable uncertainty using reinforcement learning. In Proceedings of the 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm), Tempe, AZ, USA, 11–13 November 2020; pp. 1–6.
74. Yao, S.; Gu, J.; Zhang, H.; Wang, P.; Liu, X.; Zhao, T. Resilient load restoration in microgrids considering mobile energy storage fleets: A deep reinforcement learning approach. In Proceedings of the 2020 IEEE Power & Energy Society General Meeting (PESGM), Montreal, QC, Canada, 3–6 August 2020; pp. 1–5.
75. Bedoya, J.C.; Wang, Y.; Liu, C.C. Distribution system resilience under asynchronous information using deep reinforcement learning. *IEEE Trans. Power Syst.* **2021**, *36*, 4235–4245. [CrossRef]
76. Hosseini, M.M.; Parvania, M. Resilient operation of distribution grids using deep reinforcement learning. *IEEE Trans. Ind. Inform.* **2021**, *18*, 2100–2109. [CrossRef]
77. Du, Y.; Wu, D. Deep reinforcement learning from demonstrations to assist service restoration in islanded microgrids. *IEEE Trans. Sustain. Energy* **2022**, *13*, 1062–1072. [CrossRef]
78. Gautam, M. Distribution System Resilience Enhancement Using Movable Energy Resources. Ph.D. Thesis, University of Nevada Reno, Reno, NV, USA, 2022.
79. Dehghani, N.L.; Jeddi, A.B.; Shafieezadeh, A. Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning. *Appl. Energy* **2021**, *285*, 116355. [CrossRef]
80. Li, Q.; Zhang, X.; Guo, J.; Shan, X.; Wang, Z.; Li, Z.; Chi, K.T. Integrating reinforcement learning and optimal power dispatch to enhance power grid resilience. *IEEE Trans. Circuits Syst. II Express Briefs* **2021**, *69*, 1402–1406. [CrossRef]
81. Zhao, T.; Wang, J. Learning sequential distribution system restoration via graph-reinforcement learning. *IEEE Trans. Power Syst.* **2021**, *37*, 1601–1611. [CrossRef]
82. Wang, Y.; Qiu, D.; Strbac, G. Multi-agent deep reinforcement learning for resilience-driven routing and scheduling of mobile energy storage systems. *Appl. Energy* **2022**, *310*, 118575. [CrossRef]
83. Qiu, D.; Wang, Y.; Zhang, T.; Sun, M.; Strbac, G. Hierarchical multi-agent reinforcement learning for repair crews dispatch control towards multi-energy microgrid resilience. *Appl. Energy* **2023**, *336*, 120826. [CrossRef]
84. Nie, H.; Chen, Y.; Xia, Y.; Huang, S.; Liu, B. Optimizing the post-disaster control of islanded microgrid: A multi-agent deep reinforcement learning approach. *IEEE Access* **2020**, *8*, 153455–153469. [CrossRef]
85. Abdelmalak, M.; Hosseinpour, H.; Hotchkiss, E.; Ben-Idris, M. Post-Disaster Generation Dispatching for Enhanced Resilience: A Multi-Agent Deep Deterministic Policy Gradient Learning Approach. In Proceedings of the 2022 North American Power Symposium (NAPS), Salt Lake City, UT, USA, 9–11 October 2022 ; pp. 1–6.
86. Gautam, M.; Abdelmalak, M.; Ben-Idris, M.; Hotchkiss, E. Post-Disaster Microgrid Formation for Enhanced Distribution System Resilience. In Proceedings of the 2022 Resilience Week (RWS), National Harbor, MD, USA, 26–29 September 2022; pp. 1–6.
87. Gautam, M.; Abdelmalak, M.; MansourLakouraj, M.; Benidris, M.; Livani, H. Reconfiguration of distribution networks for resilience enhancement: A deep reinforcement learning-based approach. In Proceedings of the 2022 IEEE Industry Applications Society Annual Meeting (IAS), Detroit, MI, USA, 9–14 October 2022; pp. 1–6.
88. Wang, Y.; Qiu, D.; Teng, F.; Strbac, G. Towards microgrid resilience enhancement via mobile power sources and repair crews: A multi-agent reinforcement learning approach. *IEEE Trans. Power Syst.* 2023, *early access*. [CrossRef]
89. Ahrens, M.; Kern, F.; Schmeck, H. Strategies for an adaptive control system to improve power grid resilience with smart buildings. *Energies* **2021**, *14*, 4472. [CrossRef]
90. Deshpande, K.; Möhl, P.; Hämmerle, A.; Weichhart, G.; Zörrer, H.; Pichler, A. Energy Management Simulation with Multi-Agent Reinforcement Learning: An Approach to Achieve Reliability and Resilience. *Energies* **2022**, *15*, 7381. [CrossRef]
91. Zhang, T.; Sun, M.; Qiu, D.; Zhang, X.; Strbac, G.; Kang, C. A Bayesian Deep Reinforcement Learning-based Resilient Control for Multi-Energy Micro-gird. *IEEE Trans. Power Syst.* **2023**, *38*, 5057–5072. [CrossRef]
92. Wang, B.; Li, Y.; Ming, W.; Wang, S. Deep reinforcement learning method for demand response management of interruptible load. *IEEE Trans. Smart Grid* **2020**, *11*, 3146–3155. [CrossRef]
93. Raman, N.S.; Gaikwad, N.; Barooah, P.; Meyn, S.P. Reinforcement learning-based home energy management system for resiliency. In Proceedings of the 2021 American Control Conference (ACC), New Orleans, LA, USA, 25–28 May 2021; pp. 1358–1364.
94. Tightiz, L.; Yang, H. Resilience microgrid as power system integrity protection scheme element with reinforcement learning based management. *IEEE Access* **2021**, *9*, 83963–83975. [CrossRef]

95. Hasan, M.M.; Zaman, I.; He, M.; Giesselmann, M. Reinforcement Learning-Based Control for Resilient Community Microgrid Applications. *J. Power Energy Eng.* **2022**, *10*, 1–13. [CrossRef]
96. Bhusal, N.; Gautam, M.; Benidris, M. Detection of cyber attacks on voltage regulation in distribution systems using machine learning. *IEEE Access* **2021**, *9*, 40402–40416. [CrossRef]
97. Mehrdad, S.; Mousavian, S.; Madraki, G.; Dvorkin, Y. Cyber-physical resilience of electrical power systems against malicious attacks: A review. *Curr. Sustain. Energy Rep.* **2018**, *5*, 14–22. [CrossRef]
98. Elsayed, M.; Erol-Kantarci, M.; Kantarci, B.; Wu, L.; Li, J. Low-latency communications for community resilience microgrids: A reinforcement learning approach. *IEEE Trans. Smart Grid* **2019**, *11*, 1091–1099. [CrossRef]
99. Zhang, M.; Wu, Z.; Yan, J.; Lu, R.; Guan, X. Attack-resilient optimal PMU placement via reinforcement learning guided tree search in smart grids. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 1919–1929. [CrossRef]
100. Wei, F.; Wan, Z.; He, H. Cyber-attack recovery strategy for smart grid based on deep reinforcement learning. *IEEE Trans. Smart Grid* **2019**, *11*, 2476–2486. [CrossRef]
101. Zhang, H.; Yue, D.; Dou, C.; Xie, X.; Li, K.; Hancke, G.P. Resilient Optimal Defensive Strategy of TSK Fuzzy-Model-Based Microgrids' System via a Novel Reinforcement Learning Approach. *IEEE Trans. Neural Netw. Learn. Syst.* **2021,** *34*, 1921–1931. [CrossRef]
102. Etezadifar, M.; Karimi, H.; Aghdam, A.G.; Mahseredjian, J. Resilient Event Detection Algorithm for Non-intrusive Load Monitoring under Non-ideal Conditions using Reinforcement Learning. *IEEE Trans. Ind. Appl.* 2023, *early access*. [CrossRef]
103. Zhang, H.; Yue, D.; Dou, C.; Hancke, G.P. Resilient optimal defensive strategy of micro-grids system via distributed deep reinforcement learning approach against FDI attack. *IEEE Trans. Neural Netw. Learn. Syst.* 2022, *early access*. [CrossRef]
104. Sahu, A.; Venkatraman, V.; Macwan, R. Reinforcement Learning Environment for Cyber-Resilient Power Distribution System. *IEEE Access* **2023**, *11*, 127216–127228. [CrossRef]
105. Zeng, L.; Qiu, D.; Sun, M. Resilience enhancement of multi-agent reinforcement learning-based demand response against adversarial attacks. *Appl. Energy* **2022**, *324*, 119688. [CrossRef]
106. Fard, N.E.; Selmic, R.R. Data Transmission Resilience to Cyber-attacks on Heterogeneous Multi-agent Deep Reinforcement Learning Systems. In Proceedings of the 2022 17th International Conference on Control, Automation, Robotics and Vision (ICARCV), Singapore, 11–13 December 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 758–764.
107. Chen, P.; Liu, S.; Chen, B.; Yu, L. Multi-agent reinforcement learning for decentralized resilient secondary control of energy storage systems against DoS attacks. *IEEE Trans. Smart Grid* **2022**, *13*, 1739–1750. [CrossRef]
108. Guo, Y.; Wang, L.; Liu, Z.; Shen, Y. Reinforcement-learning-based dynamic defense strategy of multistage game against dynamic load altering attack. *Int. J. Electr. Power Energy Syst.* **2021**, *131*, 107113. [CrossRef]
109. Huang, Y.; Huang, L.; Zhu, Q. Reinforcement learning for feedback-enabled cyber resilience. *Annu. Rev. Control* **2022**, *53*, 273–295. [CrossRef]
110. Ma, S.; Chen, B.; Wang, Z. Resilience enhancement strategy for distribution systems under extreme weather events. *IEEE Trans. Smart Grid* **2016**, *9*, 1442–1451. [CrossRef]
111. Pang, K.; Zhou, J.; Tsianikas, S.; Ma, Y. Deep Reinforcement Learning Based Microgrid Expansion Planning with Battery Degradation and Resilience Enhancement. In Proceedings of the 2021 3rd International Conference on System Reliability and Safety Engineering (SRSE), Harbin, China, 26–28 November 2021; pp. 251–257.
112. Pang, K.; Zhou, J.; Tsianikas, S.; Ma, Y. Deep reinforcement learning for resilient microgrid expansion planning with multiple energy resource. *Qual. Reliab. Eng. Int.* **2022**. [CrossRef]
113. Paul, S.; Dubey, A.; Poudel, S. Planning for resilient power distribution systems using risk-based quantification and Q-learning. In Proceedings of the 2021 IEEE Power & Energy Society General Meeting (PESGM), Washington, DC, USA, 26–29 July 2021; pp. 1–5.
114. Ibrahim, M.; Alsheikh, A.; Elhafiz, R. Resiliency assessment of power systems using deep reinforcement learning. *Comput. Intell. Neurosci.* **2022**, *2022*, 2017366. [CrossRef] [PubMed]
115. Abdelmoaty, A.; Naboulsi, D.; Dahman, G.; Poitau, G.; Gagnon, F. Resilient topology design for wireless backhaul: A deep reinforcement learning approach. *IEEE Wirel. Commun. Lett.* **2022**, *11*, 2532–2536. [CrossRef]
116. Nguyen, N.D.; Nguyen, T.; Nahavandi, S. Multi-agent behavioral control system using deep reinforcement learning. *Neurocomputing* **2019**, *359*, 58–68. [CrossRef]
117. Sami, H.; Mourad, A.; Otrok, H.; Bentahar, J. Demand-driven deep reinforcement learning for scalable fog and service placement. *IEEE Trans. Serv. Comput.* **2021**, *15*, 2671–2684. [CrossRef]
118. Qu, G.; Lin, Y.; Wierman, A.; Li, N. Scalable multi-agent reinforcement learning for networked systems with average reward. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2074–2086.
119. Penmetcha, M.; Min, B.C. A deep reinforcement learning-based dynamic computational offloading method for cloud robotics. *IEEE Access* **2021**, *9*, 60265–60279. [CrossRef]
120. Strehl, A.L.; Li, L.; Wiewiora, E.; Langford, J.; Littman, M.L. PAC model-free reinforcement learning. In Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 881–888.
121. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter, F.; Wang, J.; Yang, Y.; Knoll, A. A review of safe reinforcement learning: Methods, theory and applications. *arXiv* **2022**, arXiv:2205.10330.

122. Cobbe, K.; Klimov, O.; Hesse, C.; Kim, T.; Schulman, J. Quantifying generalization in reinforcement learning. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 1282–1289.
123. Packer, C.; Gao, K.; Kos, J.; Krähenbühl, P.; Koltun, V.; Song, D. Assessing generalization in deep reinforcement learning. *arXiv* **2018**, arXiv:1810.12282.
124. Dann, C.; Li, L.; Wei, W.; Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 1507–1516.
125. Yang, J.; Soltan, A.A.; Eyre, D.W.; Clifton, D.A. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. *Nat. Mach. Intell.* **2023**, *5*, 884–894. [CrossRef]
126. Abel, D.; MacGlashan, J.; Littman, M.L. Reinforcement Learning as a Framework for Ethical Decision Making. In *AAAI Workshop: AI, Ethics, and Society*; AAAI Press: Palo Alto, CA, USA, 2016; Volume 16, p. 2.
127. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1928–1937.
128. Narvekar, S.; Peng, B.; Leonetti, M.; Sinapov, J.; Taylor, M.E.; Stone, P. Curriculum learning for reinforcement learning domains: A framework and survey. *J. Mach. Learn. Res.* **2020**, *21*, 7382–7431.
129. Zhu, Z.; Lin, K.; Jain, A.K.; Zhou, J. Transfer learning in deep reinforcement learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13344–13362. [CrossRef]
130. Schweighofer, N.; Doya, K. Meta-learning in reinforcement learning. *Neural Netw.* **2003**, *16*, 5–9. [CrossRef]
131. Wells, L.; Bednarz, T. Explainable ai and reinforcement learning—A systematic review of current approaches and trends. *Front. Artif. Intell.* **2021**, *4*, 550030. [CrossRef]
132. Zhang, K.; Zhang, J.; Xu, P.D.; Gao, T.; Gao, D.W. Explainable AI in deep reinforcement learning models for power system emergency control. *IEEE Trans. Comput. Soc. Syst.* **2021**, *9*, 419–427. [CrossRef]
133. MansourLakouraj, M.; Gautam, M.; Livani, H.; Benidris, M. Multi-Stage Volt/VAR Support in Distribution Grids: Risk-Aware Scheduling with Real-Time Reinforcement Learning Control. *IEEE Access* **2023**, *11*, 54822–54838. [CrossRef]
134. Abel, D.; Salvatier, J.; Stuhlmüller, A.; Evans, O. Agent-agnostic human-in-the-loop reinforcement learning. *arXiv* **2017**, arXiv:1701.04079.
135. Luo, B.; Wu, Z.; Zhou, F.; Wang, B.C. Human-in-the-Loop Reinforcement Learning in Continuous-Action Space. *IEEE Trans. Neural Netw. Learn. Syst.* 2023, *early access*. [CrossRef]
136. Powers, S.; Xing, E.; Kolve, E.; Mottaghi, R.; Gupta, A. Cora: Benchmarks, baselines, and metrics as a platform for continual reinforcement learning agents. In Proceedings of the Conference on Lifelong Learning Agents, Montreal, QC, Canada, 22–24 August 2022; pp. 705–743.
137. Wu, T.; Scaglione, A.; Arnold, D. Reinforcement Learning using Physics Inspired Graph Convolutional Neural Networks. In Proceedings of the 2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 28–30 September 2022; pp. 1–8.
138. Garcıa, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.