Explaining Reinforcement Learning Policies for Power Grid Operations

Luca Marzari[1], Francesco Leofante[2] and Enrico Marchesini[3]

[1]University of Verona, Verona, Italy
[2]Imperial College London, London, United Kingdom
[2]Massachusetts Institute of Technology, Cambridge (MA), United States

Abstract

Reinforcement learning (RL) offers significant potential for improving decision-making in power grid operations by enabling adaptive and scalable control through interaction with these complex systems. However, real-world deployment of RL in this domain faces key challenges, including uncertainty in system dynamics, the need to achieve long-term objectives, and strict physical and safety constraints. Moreover, the black-box nature of deep RL models limits interpretability, making them difficult to trust their deployment in safety-critical power grid applications. Overcoming these obstacles requires close collaboration with system operators to develop RL methods that are not only effective but also transparent and reliable. In this work, we present our recent advances in applying RL to power grids and highlight the importance of combining RL algorithms with explainable artificial intelligence techniques to enable safe, interpretable, and trustworthy control solutions for power grid operations.

Keywords

Explainable AI, Reinforcement Learning, Power Grid

1. Introduction

Power grid operations are undergoing rapid transformation to support global decarbonization goals. This transition demands greater operational flexibility, enhanced reliability, and large-scale integration of variable renewable energy (VRE) sources. One key strategy enabling this shift and highlighted by transmission system operators (TSOs) is topology optimization—a cost-effective control method that dynamically reconfigures grid connectivity to alleviate congestion, handle contingencies (i.e., unexpected disruptions), and improve overall system security [1]. Another category of actions involves modifying power flows by redispatching or curtailing the output of fossil and renewable generators. These modifications are often costly, as they disrupt third-party operations and may incur additional costs. However, traditional optimization solvers based on these interventions often struggle to manage the growing variability introduced by VRE [2].

Reinforcement learning (RL) is gaining traction as a promising approach to automate real-time control in power systems. It has shown strong performance in complex, sequential decision-making tasks across domains such as games, robotics, and physics-based environments [3, 4, 5]. Despite these advances, several fundamental challenges continue to limit the real-world deployment of RL—such as managing complex system dynamics and inherent uncertainty, achieving long-term objectives, and adhering to strict physical constraints [6]. Power grids exemplify many of these issues, which remain open problems in the RL community. As such, studying realistic power grid tasks through the lens of RL presents a unique opportunity to drive progress in both critical infrastructure management and RL research. Yet, development in this area is slowed by the absence of standardized benchmarks that can guide progress and generate actionable insights for tackling real-world problems.

Our current work, `RL2Grid` [7], introduces the first reinforcement learning benchmark specifically tailored to realistic power grid operations, developed in collaboration with leading transmission system

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23–24, 2025, Trieste, Italy
luca.marzari@univr.it (L. Marzari); f.leofante@imperial.ac.uk (F. Leofante); emarche@mit.edu (E. Marchesini)
https://lmarza.github.io (L. Marzari); https://fraleo.github.io (F. Leofante); https://emarche.github.io (E. Marchesini)
0000-0002-0069-0182 (L. Marzari); 0000-0001-8245-9429 (F. Leofante); 0000-0003-1858-7279 (E. Marchesini)

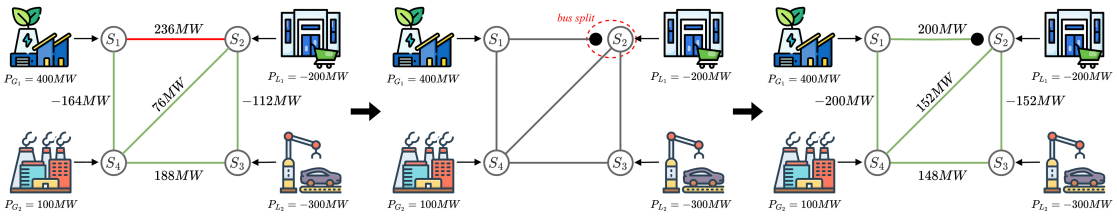

Figure 1: Toy example of a power grid where a bus split discrete topological action (center) addresses an overloaded line (depicted in red on the left).

operators (TSOs). The benchmark is designed to drive progress in grid control and advance the development of RL methods by offering a standardized suite of increasingly complex tasks. These tasks reflect the real-world challenges of power grid management, including the combinatorial complexity of the action space typical in grid operations. Figure 1 depicts a simplified power grid setup with four substations connected by transmission lines (edges), where each substation contains buses linked to two generators and two loads. Generators supply electricity to meet demand, and power is transmitted across the network, incurring losses due to line resistance. Substations, each comprising multiple buses, can partially control power routing by switching connections. Every component in this system is subject to physical constraints: generators have ramp rate limits that restrict sudden changes in output, and transmission lines have thermal limits, where sustained overloads can lead to damage or forced disconnection. Thus, advancing RL algorithms on top of `RL2Grid` to tackle the unique challenges of power grid operations has the potential to significantly benefit both RL research and deployment.

However, successfully deploying RL in these complex physical systems requires close collaboration with system operators and the development of solutions that are not only effective but also safe, transparent, and trustworthy. Achieving this is particularly difficult due to the black-box nature of deep neural networks, which underpin most scalable RL methods. Regarding safety, our ongoing research on formal and probabilistic verification for deep neural networks (FV) [8, 9, 10, 11, 12] led us to design novel safe RL algorithms, detailed in Section 3. On top of this, we aim to make RL models output decisions that are intelligible (transparent) and thus acceptable by system operators, making explainability a critical requirement for deployment. Our future research, detailed in Section 4, will focus on explainable AI (XAI) techniques tailored to RL, with the goal of improving the interpretability of learned policies [13, 14, 15].

2. RL for Power Grid Operations

We model power grid operation tasks as Markov decision processes (MDPs)—a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \rho, R, \gamma)$; $\mathcal{S}$ and $\mathcal{A}$ are the finite sets of states and actions, respectively, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability distribution, $\rho : \mathcal{S} \to [0, 1]$ is the initial uniform state distribution, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function, and $\gamma \in [0, 1)$ is the discount factor. In policy optimization algorithms [16], agents learn a parameterized stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$, modeling the probability to take an action $a_t \in \mathcal{A}$ in a state $s_t \in \mathcal{S}$ at a certain step t. The agent gets a reward for its actions, and the goal is to find the parameters that maximize the expected discounted reward $\psi(\pi) = \mathbb{E}_{\tau\sim\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)]$, where $\tau := (s_0, a_0, s_1, a_1, \dots)$ is a trajectory with $s_0 \sim \rho(s_0)$, $a_t \sim \pi(a_t|s_t)$, $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$.

Problem setup. Following this standard MDP formalization, our RL benchmark addresses power grid operation through two types of *actions*:

- *Topology (discrete actions).* Electrical devices (e.g., loads, generators, batteries) are connected to one of two buses within each substation. These discrete actions involve selecting substations where a bus-split reconfiguration can help mitigate contingencies—unplanned events that disrupt normal grid behavior—by altering how components are interconnected.
- *Redispatch or curtailment (continuous actions).* These continuous actions involve modifying power

flows by redispatching fossil generators or curtailing output from renewable sources to maintain system stability.

These actions are conditioned on the state of the power grid—a features vector including active and reactive power injections, charge levels, and maximum productions.[1]

To encourage the agent to keep the grid operational for as long as possible while minimizing the (i) overloads and disconnections of transmission lines and (ii) economic costs related to redispatching, the environment returns a *reward signal* that is the combination of three weighted components:

- R_{survive}: A constant value awarded at each step to encourage sustained grid operation.
- R_{overload}: Penalizes line overloads and disconnections, and rewards available line capacity based on the difference between line flows and capacity limits. Disconnected lines incur a fixed penalty. This term is normalized to $[-1, 1]$.
- R_{cost}: Penalizes redispatching or curtailment actions based on deviations from planned dispatch schedules and energy losses. This component is normalized to $[-1, 0]$.

The overall reward is computed as a weighted sum $R = \alpha R_{\text{survive}} + \beta R_{\text{overload}} + \eta R_{\text{cost}}$, where α, β, and η are weights tuned in an initial grid search and set to $1.0, 0.5, 0.5$, respectively. The goal is thus to keep the grid operational over a long horizon (typically, a month of operations divided into 5 minute intervals where the agent executes an action).

Sample training environment. While the `RL2Grid` benchmark offers close to 100 tasks characterized by different power grids, action spaces, and customizable configurations, in this work we focus on an explanatory grid with 6 substations. Figure 2 shows this setup on which we test well-known RL baselines. The power grid includes 6 substations (blue circles), 3 loads (yellow triangles), and 4 generators (green pentagons). For simplicity, we only consider the case where the agent controls power injections and curtailments using a 6-dimensional continuous action space.

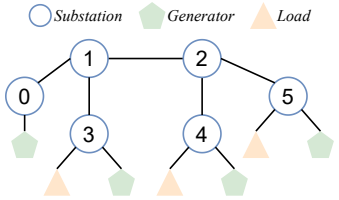

Figure 2: Explanatory power grid used in our experiments.

Data collection is performed on Xeon E5-2650 CPU nodes with 64GB of RAM, using the popular PPO and TRPO algorithms (widely-adopted baselines in the RL community [18]) as well as our recent ε version, where agents are penalized for violating operational limits (more details in the next section). To this end, we also use the cost—an auxiliary indicator function highlighting operational violations (i.e., agents get a positive cost when transmission line capacities exceed a safety limit of 95% their operational capacity).

Figure 3 reports the average return, cost, and standard error as shaded regions over 50 independent runs per method. ε-TRPO and ε-PPO are notably safer, more sample efficient, and have significantly higher performance than the baseline counterparts (TRPO and PPO).

3. Fostering Safety in RL for Power Grids

Our research naturally extends to safe RL methods [19], since ensuring safe grid operations is key to future deployments of the learned policies. In more detail, safe RL problems are typically modeled using constrained MPDs [20], where an agent (or multiple agents [21]) aims at maximizing a reward signal while limiting the accumulation of the previously mentioned cost signals under a desired threshold. However, these constraints naturally hinder exploration, failing to learn safe, effective behaviors in complex environments [22]. On top of these issues, deep neural networks (DNNs), which characterize RL policies for complex, high-dimensional tasks, are known to be vulnerable to small input variations [23, 24]. These variations can easily fool a policy to output an undesired (and unsafe) action. For these reasons, FV [8, 25] tools have arisen to tackle this issue, leveraging state-action relationships (called

[1] For the sake of clarity and brevity, we refer to RTE France [17] for an exhaustive overview of the MDP formalization, the state and action spaces, reward, and value ranges.

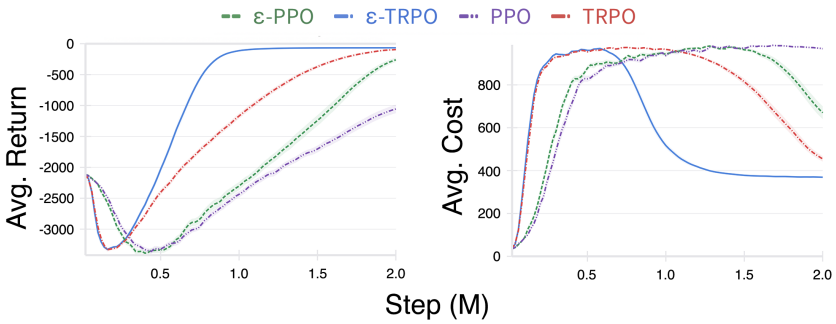

Figure 3: Comparison of ε-PPO, ε-TRPO, PPO and TRPO in the continuous redispatching and curtailment case for the explanatory power grid of Figure 2.

safety properties) to provably detect these unsafe input variations. However, applying FV during training presents many challenges. For example, the *safety properties* are hand-designed by a system designer, which may be unfeasible in complex tasks, and FV is an NP-complete problem and thus computationally untractable to apply when training RL policies [26].

We tackle the problem of fostering safety at training time by proposing a technique that collects parts of the state space where the agent is prone to unsafe actions at training time [27]. In power grid operations, and in particular in the *redispatch* task of Section 2, unsafe actions translate into critical failures due to potentially cascading failures, system instability, and generation not meeting the demand. Hence, when training our ε policies for the grid in Figure 2 we collect the state-action pairs that lead to grid instability for each unsafe action detected. Inspired by human learning, where consistently repeating tasks enhances the learning of a particular behavior, we then propose to switch between random initial training configurations for the grid state distribution, typical of RL methods, and a retraining phase in one of these collected potentially unsafe states. Crucially, to maintain the convergence properties of the original algorithm, we balance the exploration between the two scenarios using a decaying factor ε that scales down to 0 over the training.

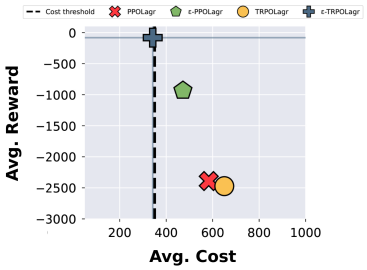

Figure 4: Pareto frontier of reward versus cost for ε-PPOLagr, ε-TRPOLagr, PPOLagr and TRPOLagr at convergence.

The results in Figure 3 already show how RL baselines augmented with this ε strategy significantly improve performance and safety. Here, we investigate whether the proposed ε approach can also improve the performance of existing safe RL methods. We summarize these additional results in Figure 4, showing the Pareto frontier of average reward versus average cost at convergence obtained by the Lagrangian implementations of PPO and TRPO (i.e., PPOLagr and TRPOLagr, respectively) as well-known safe RL algorithms [22], and their ε counterpart. These results highlight how ε retraining an agent in potentially unstable (and unsafe) power grid configurations helps agents learn to operate the grid effectively for longer periods of time when compared to existing Lagrangian algorithms. Notably, our ε retraining approach applied to existing safe RL baselines results in the best trade-off between average reward and cost, confirming the benefits of our approach in safety-critical contexts.

4. Explainability: an open challenge

Despite these advances, RL models output decisions that are not intelligible and thus acceptable by human operators, making explainability a critical requirement for deployment (see e.g., https://post.parliament.uk/research-briefings/post-pn-0735/). As a next step, we aim to co-designing RL-based controllers and XAI methods to build trust and enhance usability in critical infrastructure. In

particular, we will advance (robust) counterfactual explanations (CEs) [28, 29, 15] as an XAI technique for RL models. CEs will provide actionable insights into RL decisions through rigorous "what-if" analysis, by clarifying how changes in the input of an RL model impact system performance and by ensuring robust, interpretable explanations for real-world applications. Crucially, a recent survey highlighted that current methods for CE generation for RL are limited, thus calling for new developments in this space [13]. This interdisciplinary research bridges computational advancements with practical deployment needs, aligning RL optimization with grid operation principles and ensuring transparency and trust through robust XAI methods. The outcomes will support the development of reliable, explainable, and scalable solutions for power system operations, accelerating the transition to a fully decarbonized and resilient energy grid.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to grammar and spelling check, paraphrase, and reword. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

References

[1] B. Donnot, Grid2op-A testbed platform to model sequential decision making in power systems, https://GitHub.com/rte-france/grid2op, 2020.

[2] A. Marot, A. Kelly, M. Naglic, V. Barbesant, J. Cremer, A. Stefanov, J. Viebahn, Perspectives on future power system control centers for energy transition, Journal of Modern Power Systems and Clean Energy 10 (2022) 328–344.

[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller, Playing atari with deep reinforcement learning, in: Conference on Neural Information Processing Systems (NeurIPS), 2013.

[4] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al., Mastering the game of go with deep neural networks and tree search, Nature 529 (2016) 484–489.

[5] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs, L. Gilpin, V. Kompella, P. Khandelwal, H. Lin, P. MacAlpine, D. Oller, C. Sherstan, T. Seno, M. D. Thomure, H. Aghabozorgi, L. Barrett, R. Douglas, D. Whitehead, P. Duerr, P. Stone, M. Spranger, , H. Kitano, Outracing champion gran turismo drivers with deep reinforcement learning, Nature 62 (2022) 223–28. doi:10.1038/s41586-021-04357-7.

[6] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, T. Hester, Challenges of real-world reinforcement learning: definitions, benchmarks and analysis, Machine Learning 110 (2021) 2419–2468.

[7] E. Marchesini, B. Donnot, C. Crozier, I. Dytham, C. Merz, L. Schewe, N. Westerbeck, C. Wu, A. Marot, P. L. Donti, Rl2grid: Benchmarking reinforcement learning in power grid operations, 2025. URL: https://arxiv.org/abs/2503.23101. arXiv:2503.23101.

[8] C. Liu, T. Arnon, C. Lazarus, C. Strong, C. Barrett, M. J. Kochenderfer, et al., Algorithms for verifying deep neural networks, Foundations and Trends® in Optimization 4 (2021) 244–404.

[9] E. Marchesini, L. Marzari, A. Farinelli, C. Amato, Safe deep reinforcement learning by verifying task-level properties, in: Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, 2023, p. 1466–1475.

[10] L. Marzari, E. Marchesini, A. Farinelli, Online safety property collection and refinement for safe deep reinforcement learning in mapless navigation, in: 2023 IEEE International Conference on Robotics and Automation (ICRA), 2023, pp. 7133–7139.

[11] L. Marzari, D. Corsi, F. Cicalese, A. Farinelli, The #dnn-verification problem: counting unsafe inputs

for deep neural networks, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, 2023, pp. 217–224.
[12] L. Marzari, D. Corsi, E. Marchesini, F. Alessandro, F. Cicalese, Enumerating safe regions in deep neural networks with provable probabilistic guarantees, Proceedings of the AAAI Conference on Artificial Intelligence (2024) 21387–21394.
[13] J. Gajcin, I. Dusparic, Redefining counterfactual explanations for reinforcement learning: Overview, challenges and opportunities, ACM Computing Surveys 56 (2024) 1–33.
[14] J. Jiang, F. Leofante, A. Rago, F. Toni, Robust counterfactual explanations in machine learning: a survey, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, 2024, pp. 8086–8094.
[15] F. Leofante, M. Wicker, Robust Explainable AI, Springer Nature, 2025.
[16] E. Marchesini, C. Amato, Improving deep policy gradients with value function search, in: International Conference on Learning Representations (ICLR), 2023. URL: https://openreview.net/forum?id=6qZC7pfenQm.
[17] RTE France, Dive into grid2op sequential decision process, 2025. URL: https://grid2op.readthedocs.io/en/latest/mdp.html#some-constraints, accessed: 2025-05-8.
[18] J. Ji, J. Zhou, J. D. Borong Zhang, R. S. Xuehai Pan, W. Huang, Y. Geng, M. Liu, Y. Yang, Omnisafe: An infrastructure for accelerating safe reinforcement learning research, arXiv preprint arXiv:2305.09304 (2023).
[19] J. Garcıa, F. Fernández, A comprehensive survey on safe reinforcement learning, in: Journal of Machine Learning Research (JMLR), 2015.
[20] E. Altman, Constrained markov decision processes, in: CRC Press, 1999.
[21] A. A. Aydeniz, E. Marchesini, R. Loftin, C. Amato, K. Tumer, Safe entropic agents under team constraints, in: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, International Foundation for Autonomous Agents and Multiagent Systems, 2025, p. 2411–2413.
[22] J. Ji, B. Zhang, J. Zhou, X. Pan, W. Huang, R. Sun, Y. Geng, Y. Zhong, J. Dai, Y. Yang, Safety gymnasium: A unified safe reinforcement learning benchmark, in: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL: https://openreview.net/forum?id=WZmlxIuIGR.
[23] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).
[24] G. Amir, D. Corsi, R. Yerushalmi, L. Marzari, D. Harel, A. Farinelli, G. Katz, Verifying learning-based robotic navigation systems, in: 29th Int. Conf., TACAS 2023, Springer, 2023, pp. 607–627.
[25] T. Wei, H. Hu, L. Marzari, K. S. Yun, P. Niu, X. Luo, C. Liu, Modelverification.jl: A comprehensive toolbox for formally verifying deep neural networks, in: Computer Aided Verification - 37th International Conference, CAV 2025, volume 15932 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 395–408. URL: https://doi.org/10.1007/978-3-031-98679-6_18. doi:10.1007/978-3-031-98679-6_18.
[26] G. Katz, C. Barrett, D. L. Dill, K. Julian, M. J. Kochenderfer, Reluplex: An efficient smt solver for verifying deep neural networks, in: International conference on computer aided verification, Springer, 2017, pp. 97–117.
[27] L. Marzari, P. L. Donti, C. Liu, E. Marchesini, Improving policy optimization via ε-retrain, in: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems, 2025, p. 1464–1472.
[28] L. Marzari, F. Leofante, F. Cicalese, A. Farinelli, Rigorous probabilistic guarantees for robust counterfactual explanations, in: ECAI 2024, IOS Press, 2024, pp. 1059–1066.
[29] J. Jiang, L. Marzari, A. Purohit, F. Leofante, Robustx: Robust counterfactual explanations made easy, in: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, International Joint Conferences on Artificial Intelligence Organization, 2025, pp. 11067–11071. URL: https://doi.org/10.24963/ijcai.2025/1264. doi:10.24963/ijcai.2025/1264.