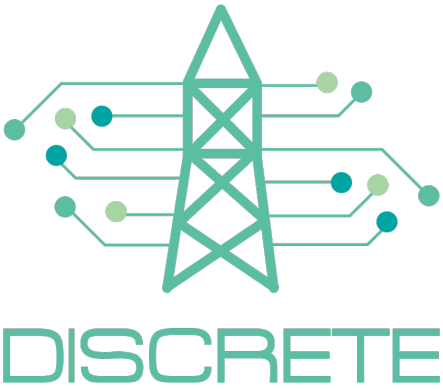

D2.3. Develop a reinforcement learning model to take autonomous decisions based on the current observable state of the electrical system

Author: Jean-François Toubeau

This deliverable is based on the publication:

C. Rasic, P. Favaro, Y. Wang and J. -F. Toubeau, (2026) "Safe Reinforcement Learning for Battery Energy Storage Participation in the Imbalance Settlement," in IEEE Transactions on Energy Markets, Policy and Regulation, ***doi: 10.1109/TEMPR.2025.3639758****.*

Table of Contents

Table of Contents 2

1. Purpose within the project 3
2. Methodological contribution 3
3. Case Study Results 4

1. Purpose within the project

This paper was carried out to investigate how reinforcement learning can be used to support autonomous operational decision-making in power systems characterized by uncertainty, partial observability, and rapidly changing operating conditions. Earlier work in DISCRETE focused on uncertainty representation, stochastic optimisation, and supervised-learning-based optimisation support. While these approaches provide powerful decision-support tools, they still rely on explicitly solving optimisation problems. This work therefore explores a complementary paradigm in which an intelligent agent learns operational policies directly from interactions with the environment and autonomously determines control actions based on the current observable state of the system.

The work addresses a key challenge for future renewable-dominated power systems: decisions often need to be made within seconds while system conditions remain uncertain and only partially observable. Reinforcement learning offers an attractive solution because once trained, the agent can react almost instantaneously without requiring the repeated solution of computationally intensive optimisation problems. Within DISCRETE, the objective is to assess whether RL can provide a realistic solution for real-time operational decision-making, capable of learning from historical data while adapting to changing operating conditions.

Aspect	Summary
Main method	Develop autonomous decision-making agents for power-system operation
Methodology	Safe Reinforcement Learning
Operational challenge	Partial observability and real-time decision making

2. Methodological contribution

The proposed framework formulates the operational decision problem as a Markov Decision Process (MDP). At each decision interval, the agent observes the current system state and determines an action that maximizes long-term operational performance. Unlike conventional optimisation approaches that require explicit models of future system behaviour, the RL agent learns directly from historical interactions and receives rewards based on the quality of its decisions. The objective is not merely to maximize immediate gains but to learn long-term strategies that anticipate future opportunities.

The selected RL architecture is based on the Soft Actor-Critic (SAC) algorithm. SAC combines policy learning and value-function estimation within an actor–critic framework while incorporating entropy regularisation to promote exploration and avoid premature convergence toward suboptimal policies. This characteristic is particularly important in energy-system applications, where profitable strategies may require short-term sacrifices in order to create future operational opportunities.

The work further introduces a safe layer. Rather than relying on reward penalties to discourage unsafe actions, the methodology incorporates a differentiable projection layer that guarantees compliance with physical system constraints. Every action proposed by the neural-network policy is projected onto the feasible operating region before implementation. This ensures that the agent remains physically feasible during both training and deployment while preserving end-to-end differentiability. The approach therefore combines the flexibility of data-driven learning with the operational reliability required for real-world deployment.

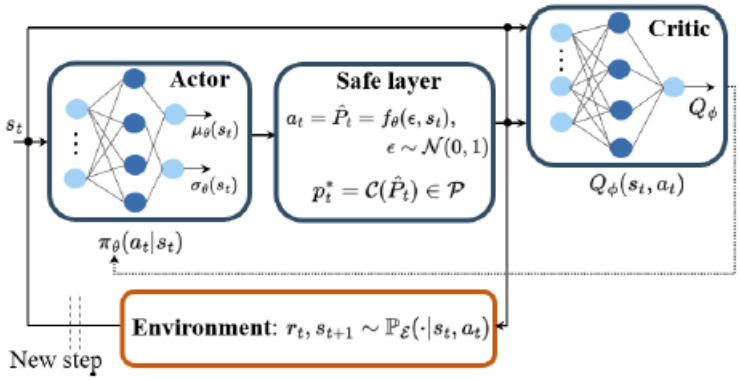

Figure - Safe Soft Actor-Critic framework.

3. Case Study Results

The results demonstrate that the proposed Safe RL framework successfully learns profitable operational strategies while maintaining full compliance with physical constraints. The differentiable projection layer eliminates infeasible actions during both training and deployment, resulting in zero observed constraint violations. This represents a significant improvement over conventional reinforcement-learning approaches, which frequently explore infeasible operating regions and may therefore generate unrealistic decisions.

The proposed SafeSAC agent also outperforms conventional RL algorithms in terms of economic performance. Compared with standard SAC and TD3 agents, the safe architecture achieves higher profitability while exhibiting more stable convergence during training. The agent learns effective long-term operational strategies directly from historical data and demonstrates strong generalisation when evaluated on unseen market conditions. Importantly for DISCRETE, decision times remain below approximately two milliseconds, making the approach suitable for real-time operational applications.

Table - Average testing performance of all decision agents over ten independent training runs. “mean agent” and “max agent” denote the models achieving, respectively, the mean and highest test profits across these runs.

	Model-based		**Safety-informed Actor-Critic**						**Actor-Critic**		**Value-based**
Metric	*Perfect foresight*	*MPC*	*SafeSAC*	*SACpen*	*SACrew*	*SafeTD3*	*TD3pen*	*TD3rew*	*SAC*	*TD3*	*DQL*
Mean agent - test profit (€/day)	20,517	4,782	8,132	7,246	7,368	8,093	7,401	7,214	6,012	5,312	3,474
Max agent - test profit (€/day)	20,517	4,782	9,283	8,346	8,789	8,933	8,312	8,291	6,880	5,924	3,917
Average training time (h)	–	–	5.47	1.72	1.73	5.18	1.71	1.71	1.72	1.71	1.98
Average infeasible actions [%]	0	0	0	2.2	2	0	2	1.8	21	27	35

The work demonstrates that reinforcement learning can serve as a viable decision-making paradigm for future power systems where rapid response and adaptability are required. Future developments beyond DISCRETE may extend the methodology toward congestion management, topology control, flexibility activation, reserve allocation, and other operational planning problems.

Overall, D2.3 contributes to DISCRETE by establishing the foundations of autonomous operational decision-making through reinforcement learning. The deliverable demonstrates that safe RL can combine computational efficiency, operational feasibility, and adaptive decision-making, thereby supporting the project’s vision of intelligent real-time operation of future renewable-dominated power systems.