RLSB

Software Engineering for Artificial Intelligence

REINFORCEMENT LEARNING SUSTAINABILITY BENCHMARK

Luca Strefezza

04 March 2025

INTRODUCTION

Context

This project addresses the energy consumption of deep reinforcement learning (DRL) solutions and their impact on the environment and business costs.

Motivation

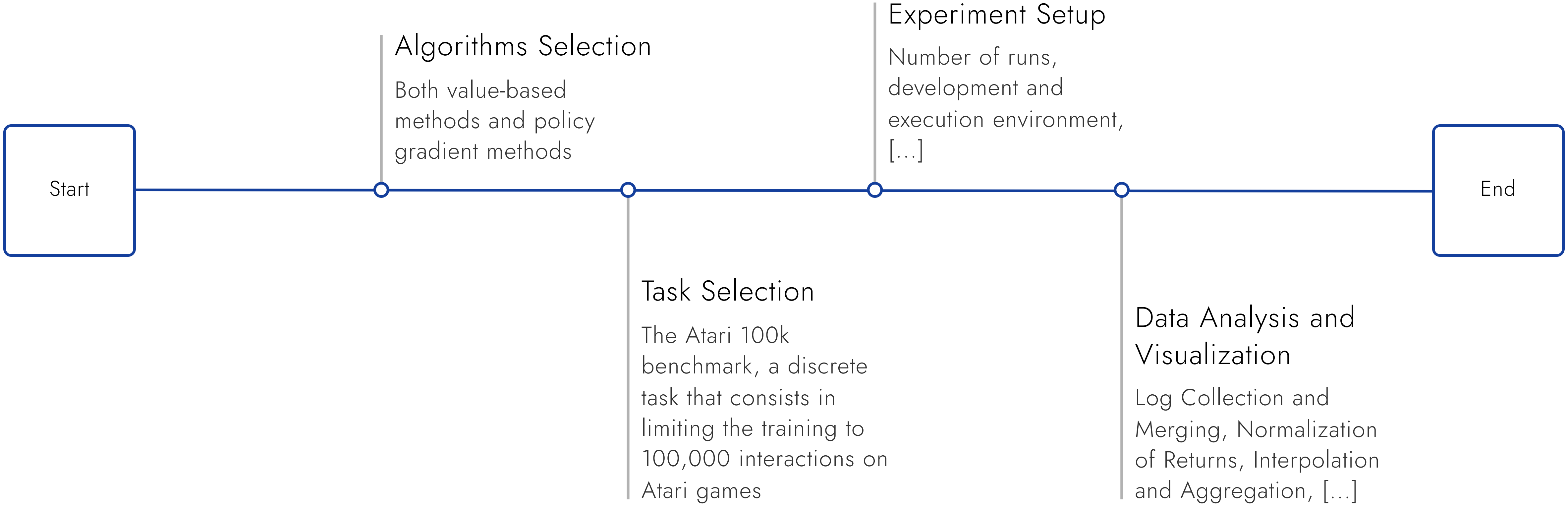
Beginning with the resurgence of the field following the development of Deep Q-Networks (DQN) numerous algorithms have been proposed: modifications to DQN or entirely different paradigms such as policy gradient methods. Although the performance of the various solutions has been extensively studied, little effort has been directed toward understanding how these modifications affect energy consumption, and how these costs compare to those of earlier methods.

Goals

This project aims to fill this gap by evaluating the trade-offs between performance and energy consumption across several widely used deep reinforcement learning algorithms.

METHODOLOGICAL STEPS

The benchmark is based on executing all algorithms for the same number of environment interactions. This allows a direct comparison of both achieved scores and energy consumption.



Start

Algorithms Selection

Both value-based methods and policy gradient methods

Task Selection

The Atari 100k benchmark, a discrete task that consists in limiting the training to 100,000 interactions on Atari games

Experiment Setup

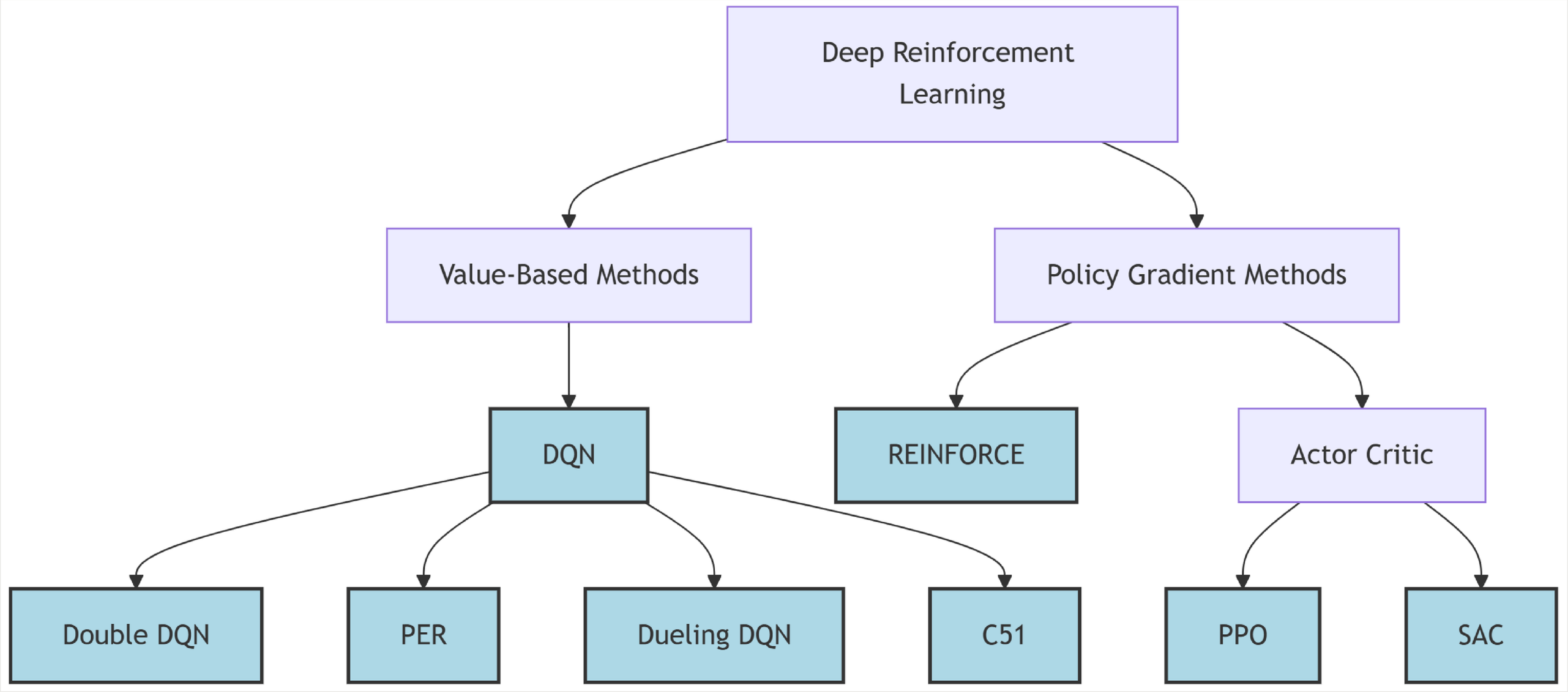
Number of runs, development and execution environment, […]

Data Analysis and Visualization

Log Collection and Merging, Normalization of Returns, Interpolation and Aggregation, […]

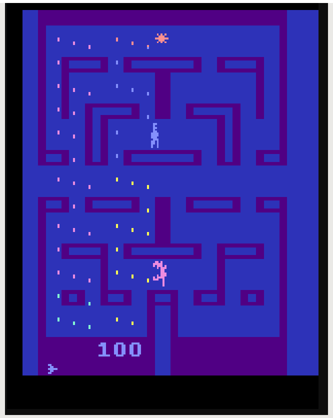
End

ALGORITHMS SELECTION

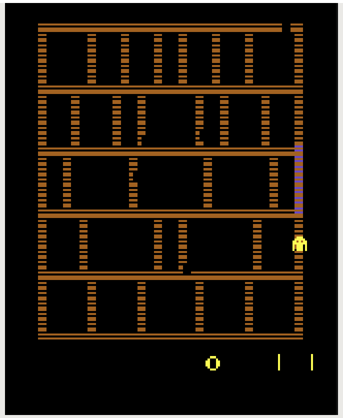


TASK SELECTION

Atari 100k is a widely used benchmark in the DRL community, well suited for evaluating the performance of almost all popular DRL algorithms, providing a range of different challenges with a discrete action space.

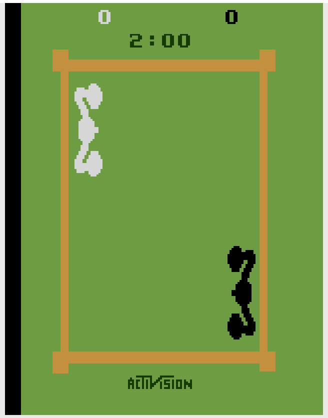


Alien

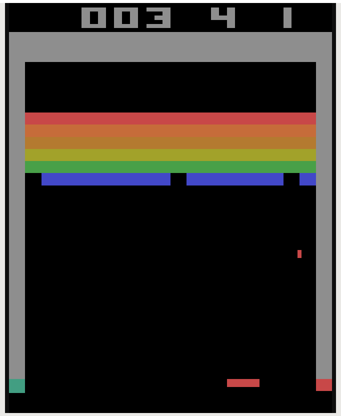


Amidar

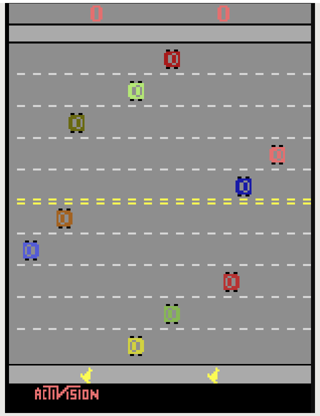
Assault



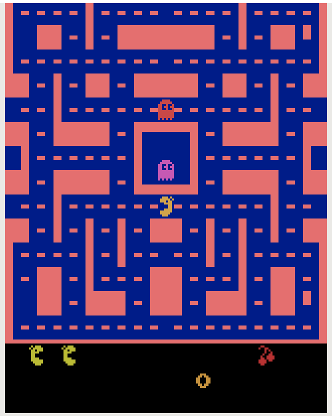
Boxing



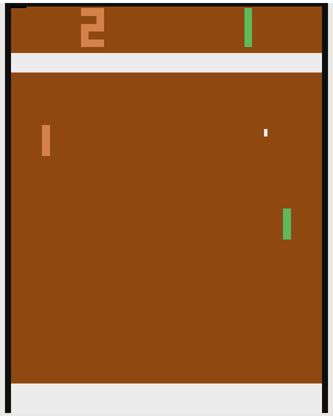
Breakout



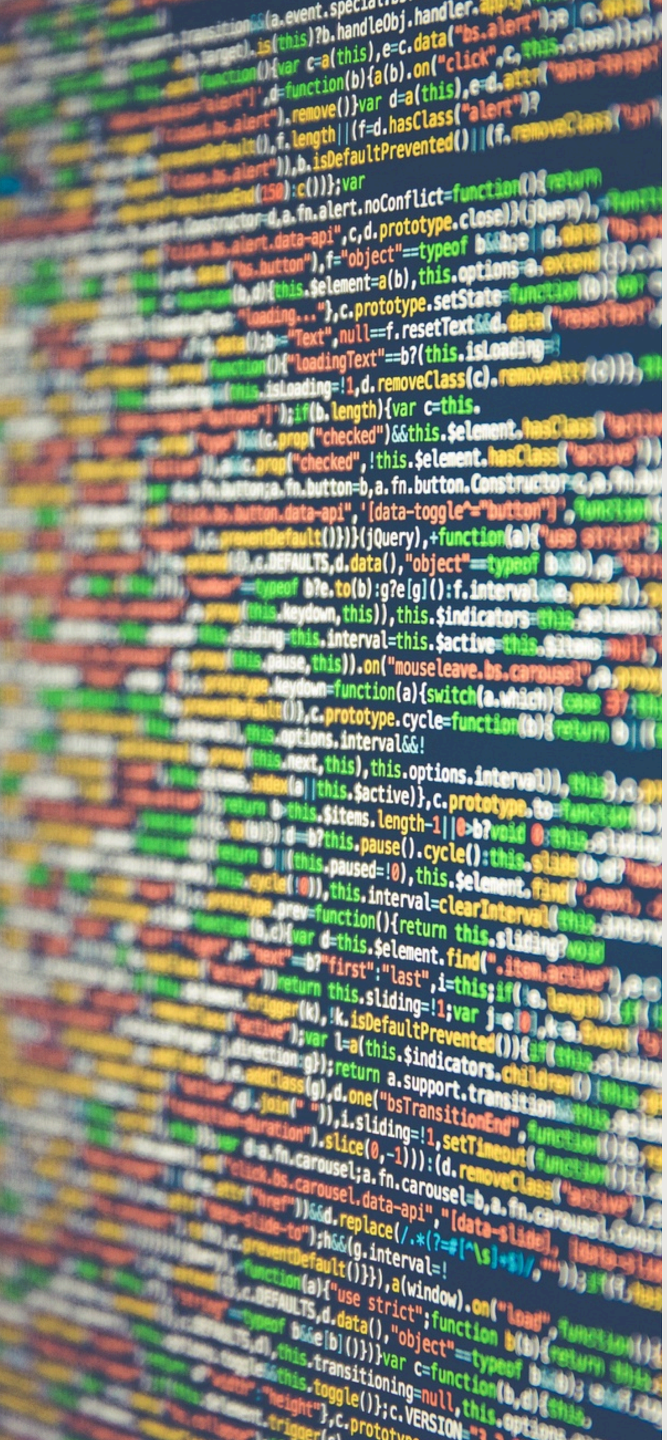
Freeway



Ms. Pac-Man



Pong



EXPERIMENT SETUP

Number of Runs

- 4 seeds to have statistically significant results (when combined with robust metrics like IQM)
- 8 environments X 4 seeds = 32 runs per algorithm.

Data Logging and Storage

1. TensorBoard
2. Weights and Biases
3. CodeCarbon

Development and Execution Environment

- Intel(R) Core(TM) i5-11400F
- GPU: NVIDIA GeForce GTX 1050 Ti
- CPU: Intel(R) Core(TM) i9-10980XE
- GPU: NVIDIA RTX A5000
- RAM: 64GB
- CleanRL

Atari Environment Configuration

Follows all the best practices (environment version NoFrameskip-v4):

1. Preprocessing and Standardization
2. Frame skipping
3. Random no-op initialization
4. Reward clipping

(Hyper)Parameter Configurations

- Primarily based on the configurations used in the original papers introducing each method.
- (Coarse) Grid Search
- Adaptation of the hyperparameters strictly connected to the number of environment interactions.

Evaluation

- Human normalization: x_norm = (x−x_random)/(x_human−x_random)
- Min-max normalization: x_norm = (x−x_min)/(x_max−x_min)
- Basic statistics computed on the normalized data
- Interquartile mean
- 10 evaluation episodes at the end of every run

DATA ANALYSIS AND VISUALIZATION

1. **Log Collection and Merging**
2. **Normalization of Returns**
3. **Interpolation and Aggregation**
4. **Plot Generation and CSV Output**

Interpolation and Aggregation

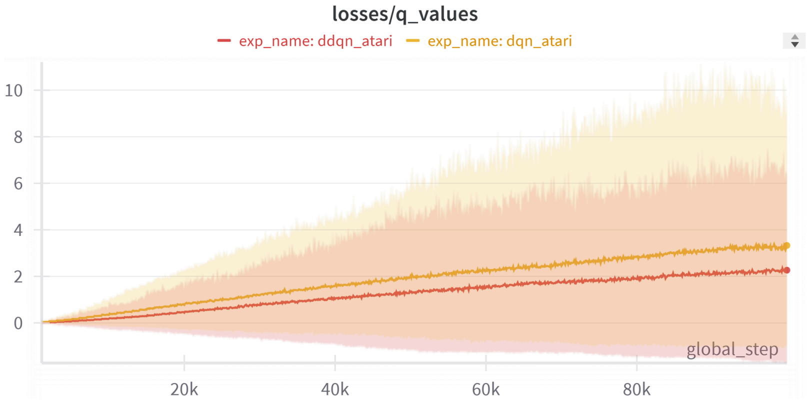
To create a consistent x-axis for plotting episodic returns, we interpolated the log of each run, sampled every 100 steps, and plotted the aggregation with 1000 data points, overcoming the issue of variable episode lengths.

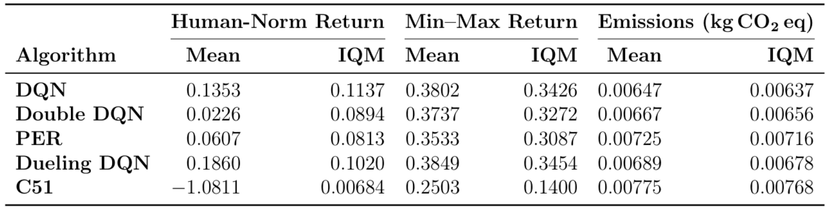
Section 2

RESULTS

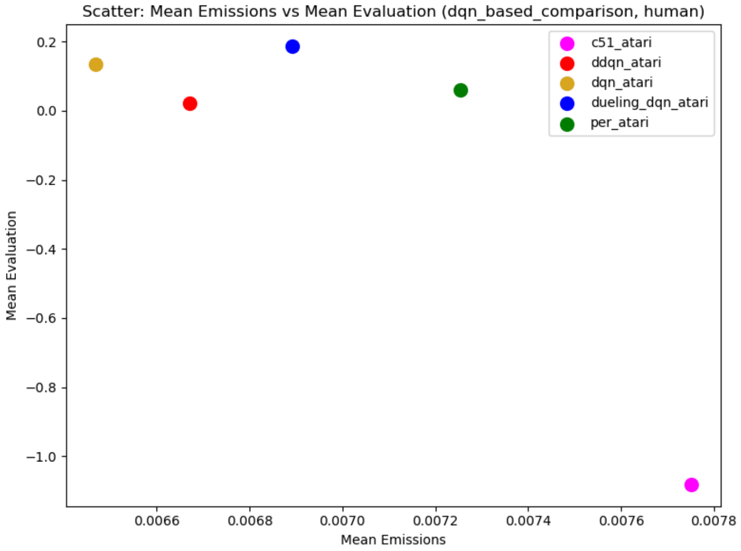
DQN BASED ALGORITHMS

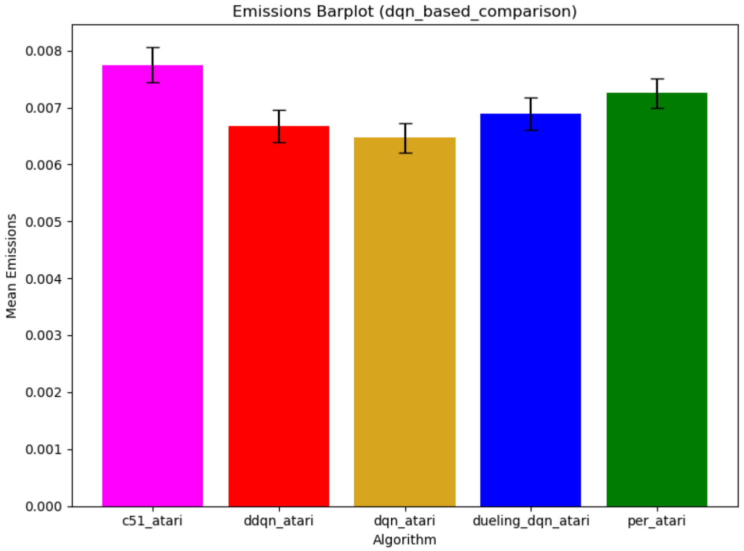
- *DQN:* Baseline performance with moderate returns and emissions around 0.006 - 0.007 kgCO₂
- *Double DQN:* Improves stability by reducing overestimation, with similar emission levels, but lacks in performance.
- *PER:* Shows variable returns with occasional negative outliers; energy consumption higher but remains comparable.
- *Dueling DQN:* Achieves higher human—normalized performance but exhibits high variance across games.
- *C51:* Tends to record higher emissions (≈0.007+ kgCO₂ eq) with performance sensitive to game dynamics.

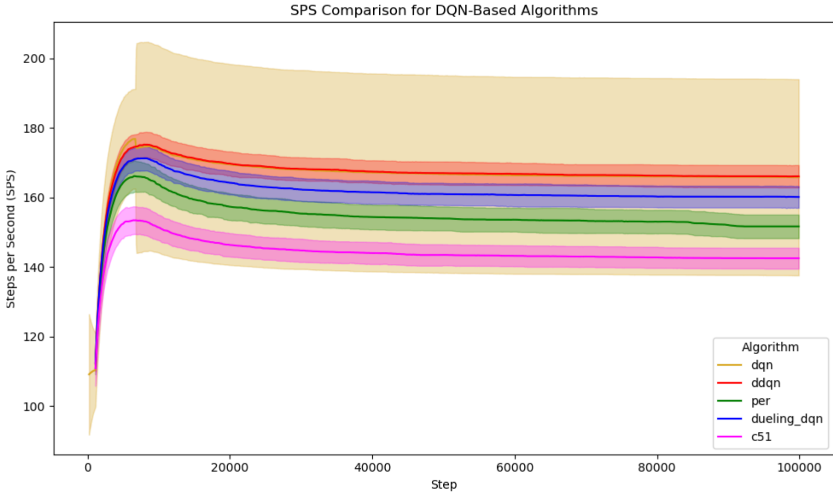




Algorithm	Human-Norm Return Mean	Human-Norm Return IQM	Min–Max Return Mean	Min–Max Return IQM	Emissions (kg CO_2 eq) Mean	Emissions (kg CO_2 eq) IQM
DQN	0.1353	0.1137	0.3802	0.3426	0.00647	0.00637
Double DQN	0.0226	0.0894	0.3737	0.3272	0.00667	0.00656
PER	0.0607	0.0813	0.3533	0.3087	0.00725	0.00716
Dueling DQN	0.1860	0.1020	0.3849	0.3454	0.00689	0.00678
C51	−1.0811	0.00684	0.2503	0.1400	0.00775	0.00768

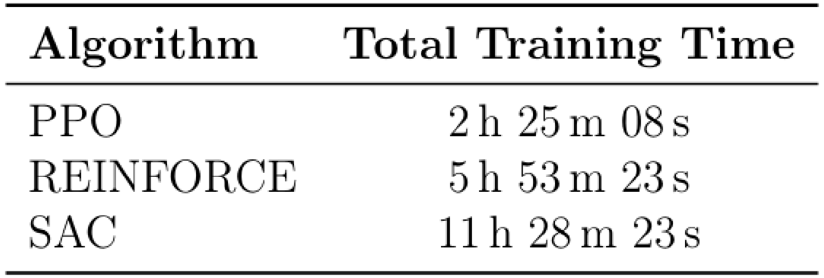




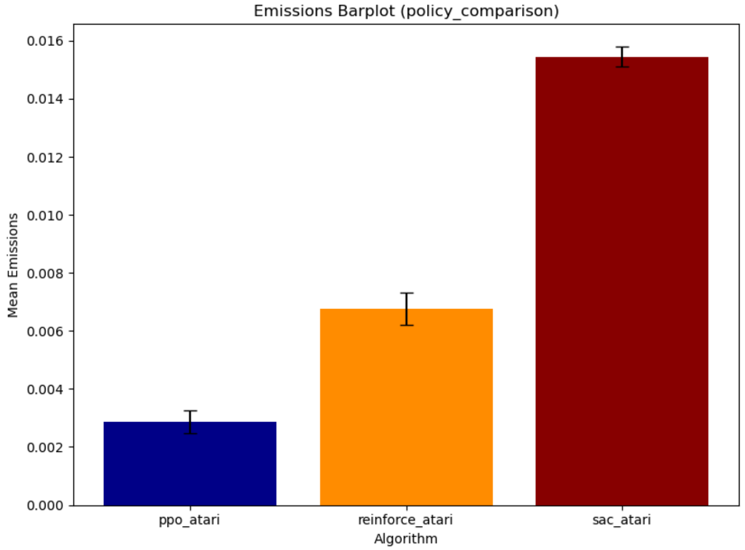


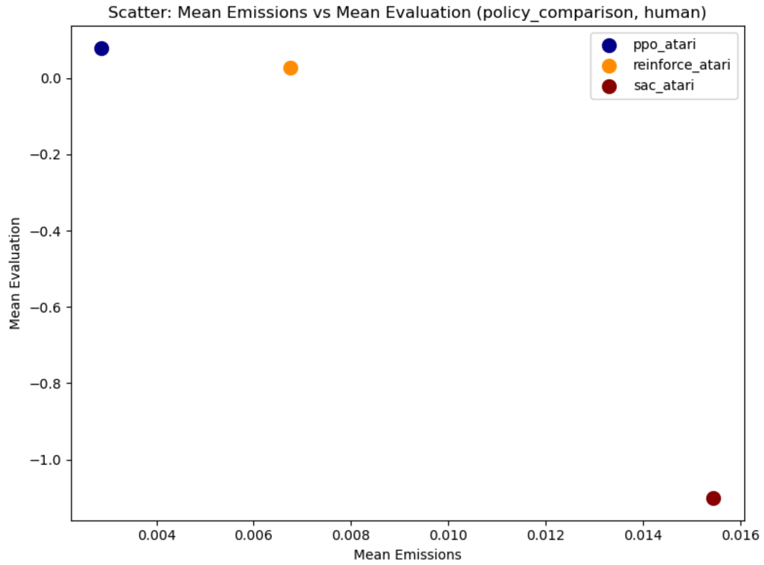
POLICY GRADIENT ALGORITHMS

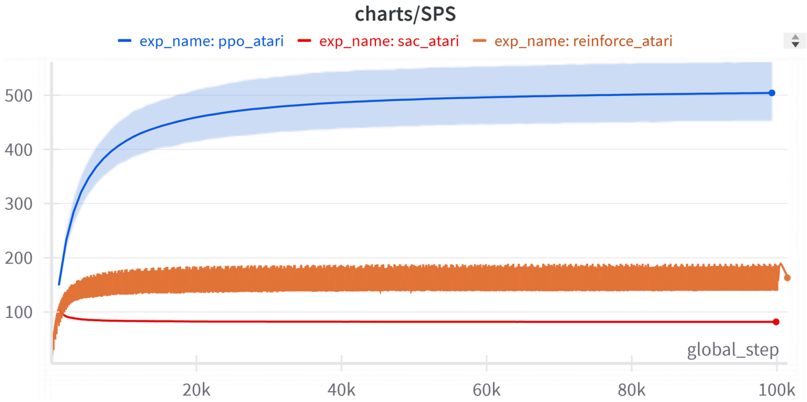
- *REINFORCE*: Displays high variance in episodic returns across seeds; moderate energy consumption.
- *PPO*: Achieves competitive performance with the lowest emissions (~0.0029 kgCO_2 eq), indicating high energy efficiency.
- *SAC*: Offers strong performance in select environments, but records the highest energy consumption (~0.015 kgCO_2 eq) and slower throughput.



Algorithm	Total Training Time
PPO	2 h 25 m 08 s
REINFORCE	5 h 53 m 23 s
SAC	11 h 28 m 23 s





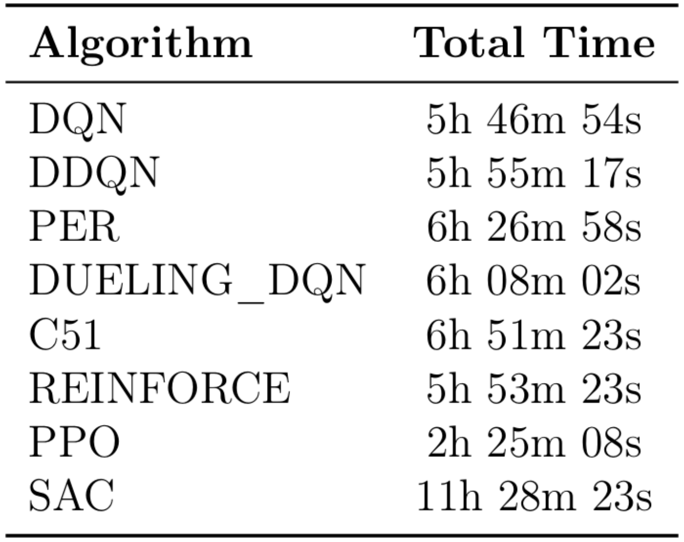


OVERALL ALGORITHM COMPARISON

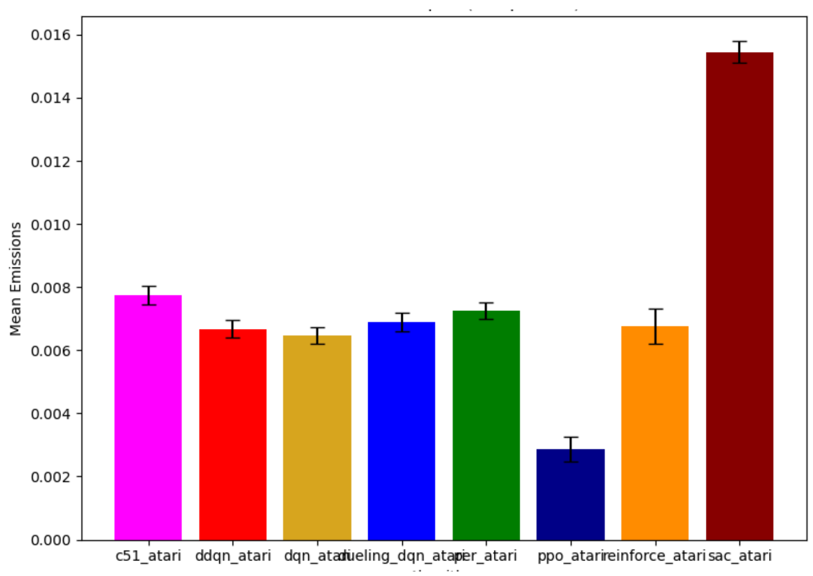
- *Performance*:
  DQN—based methods deliver consistent, moderate returns, while policy gradient methods show higher variability

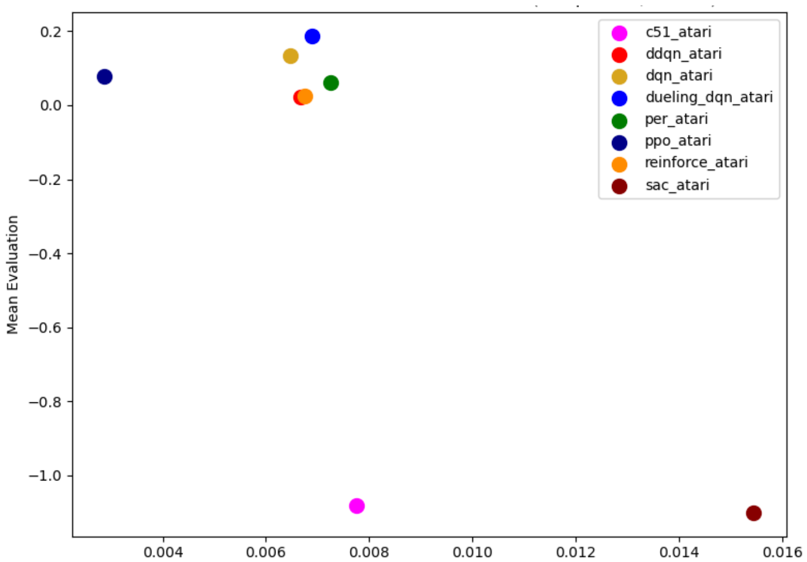
- *Energy Consumption*:
  PPO stands out with low emissions, whereas SAC incurs the highest cost

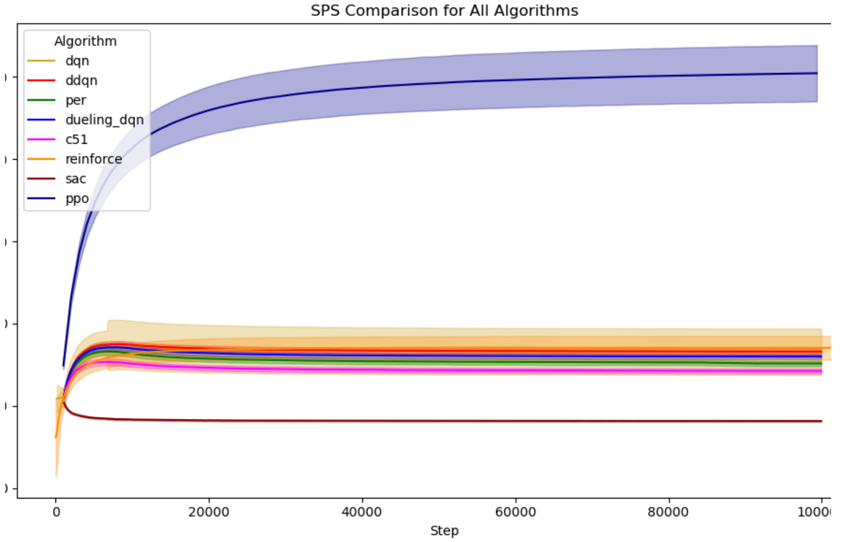
- *Runtime & Throughput*:
  DQN variants generally run faster (higher SPS), with PPO leveraging parallelism for efficiency



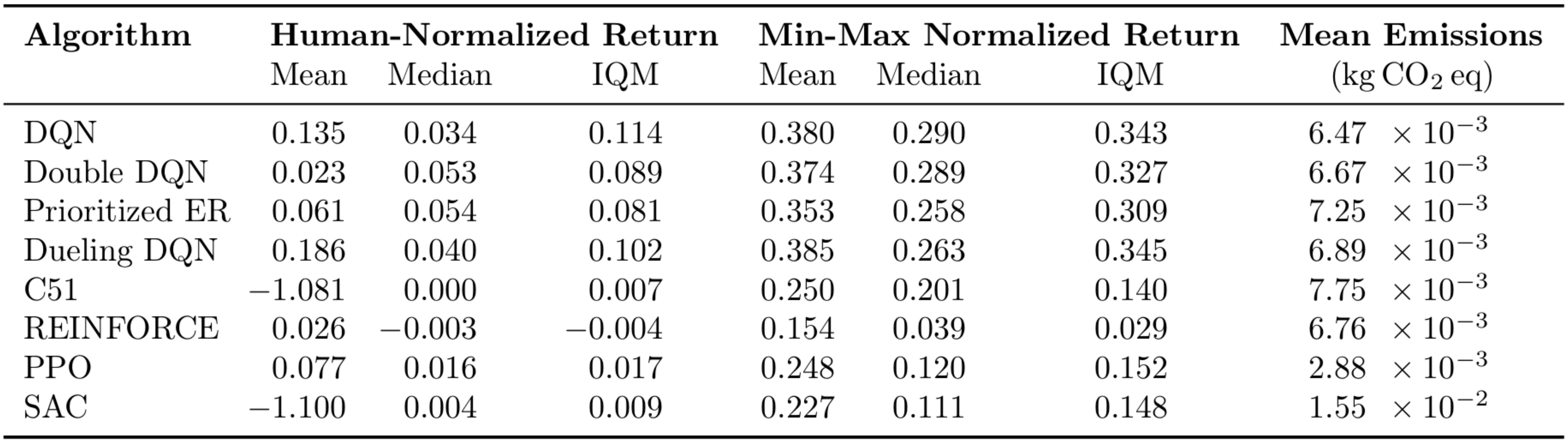
Algorithm	Total Time
DQN	5h 46m 54s
DDQN	5h 55m 17s
PER	6h 26m 58s
DUELING_DQN	6h 08m 02s
C51	6h 51m 23s
REINFORCE	5h 53m 23s
PPO	2h 25m 08s
SAC	11h 28m 23s





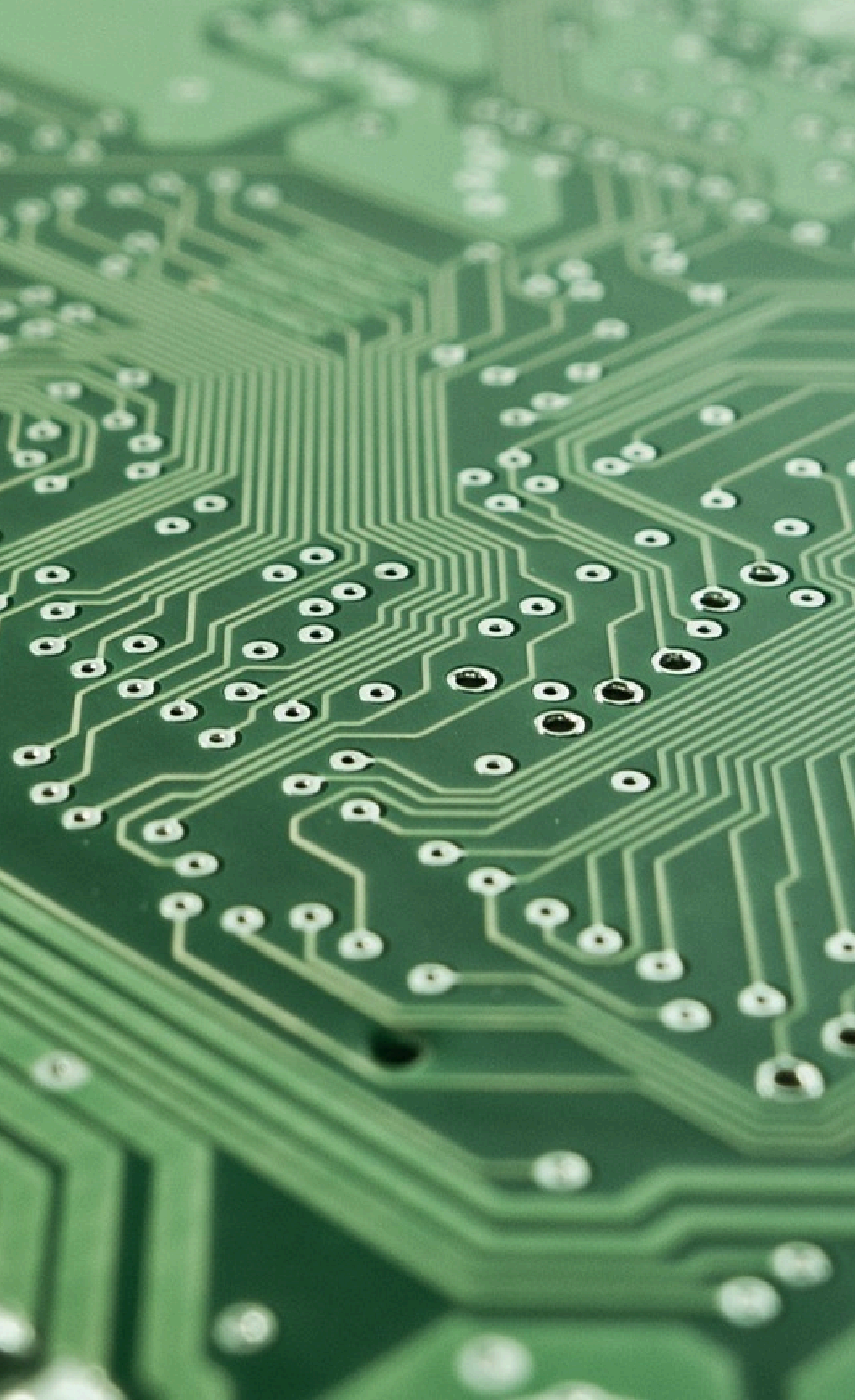


IMPLICATIONS OF THE RESULTS



Algorithm	**Human-Normalized Return**			**Min-Max Normalized Return**			**Mean Emissions**
	Mean	Median	IQM	Mean	Median	IQM	(kg CO_2 eq)
DQN	0.135	0.034	0.114	0.380	0.290	0.343	6.47×10^{-3}
Double DQN	0.023	0.053	0.089	0.374	0.289	0.327	6.67×10^{-3}
Prioritized ER	0.061	0.054	0.081	0.353	0.258	0.309	7.25×10^{-3}
Dueling DQN	0.186	0.040	0.102	0.385	0.263	0.345	6.89×10^{-3}
C51	−1.081	0.000	0.007	0.250	0.201	0.140	7.75×10^{-3}
REINFORCE	0.026	−0.003	−0.004	0.154	0.039	0.029	6.76×10^{-3}
PPO	0.077	0.016	0.017	0.248	0.120	0.152	2.88×10^{-3}
SAC	−1.100	0.004	0.009	0.227	0.111	0.148	1.55×10^{-2}

- A clear trade–off exists between performance and energy consumption
- DQN–based methods strike a balance, while policy gradient methods offer both extremes (high efficiency vs. high emissions)
- PPO’s energy efficiency demonstrates the potential for sustainable DRL in real–world applications, in particular in RLHF, RLAIF, and training for Chain of Thought



FUTURE RESEARCH DIRECTIONS

- Extending Training Horizons
- Wider Environment Coverage
- Hardware-Specific Optimization and Architecture
- RL Approaches for LLM Fine-Tuning
  - PPO is almost universally employed In LLM fine-tuning. A recent work challenges the need to use a complex algorithm like PPO and propose a simpler REINFORCE-style approach. This might suggest a reduction in emissions too, however, our benchmark reveals a counterintuitive result: despite its simplicity, REINFORCE actually produces higher emissions than PPO under the 100k-step training regime. A specific comparison could be useful.
  - Recent research after the release of DeepSeek-R1 shows that the particular RL algorithm used to induce the chain-of-thought (CoT) reasoning in LLMs makes no difference in terms of performance. This suggests that if a more sustainable RL algorithm can achieve the same chain-of-thought benefits, it would be highly attractive for practical applications, from here the need to compare PPO, GRPO, and alternative policy-gradient methods under the lens of the energy-efficiency/performance trade-off.

THANK YOU