Modern Reinforcement Learning: A Practical Guide

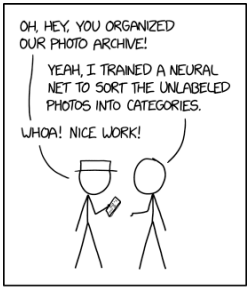
How to Use RL in 2025 with Gymnasium, PPO, and RLHF

Prof. Dr. Jörg Osterrieder

Methods and Algorithms — MSc Data Science

Spring 2026

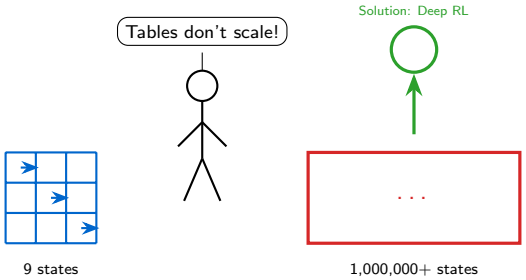
You Know RL — Now How Do People Actually Use It?



You already know what RL is (L06b). Now: how do practitioners actually **use** *it today?*

- Deep RL libraries that work out of the box
- PPO as the workhorse algorithm
- RLHF — how ChatGPT learned to be helpful

The Scale Problem: Why Tables Break Down



Classic RL stores a table of values. Modern RL uses neural networks to handle any state space.

A chess board has $\sim 10^{44}$ possible states. No table can store that. Neural networks generalize.

From Tables to Neural Networks

Classic RL (L06b)

- Q-table stores one value per state
- Works for small grids (9–100 states)
- Tabular lookup — no generalization

Modern RL (Today)

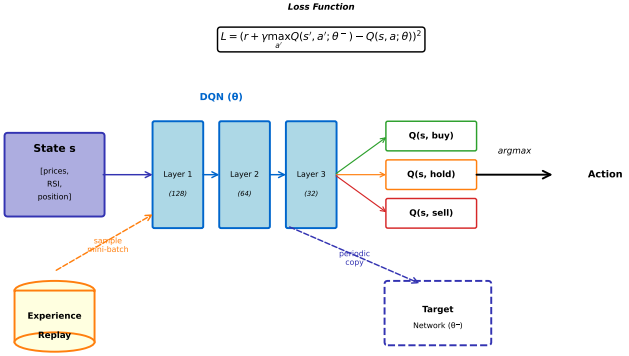
- Neural network approximates Q-values
- Works for any state space (millions+)
- Learns to generalize across states

The Key Insight

Replace the table with a neural network = Deep RL.

DQN (2015) was the breakthrough: a neural network that plays Atari games from raw pixels.

DQN: The Neural Network That Plays Atari



Deep Q-Network Architecture for Trading

- State (pixels, numbers) goes into the neural network
- Network outputs a Q-value for every possible action
- Pick the action with the highest Q-value

DQN = Deep Q-Network. DeepMind used this to beat human experts at 49 Atari games (2015).

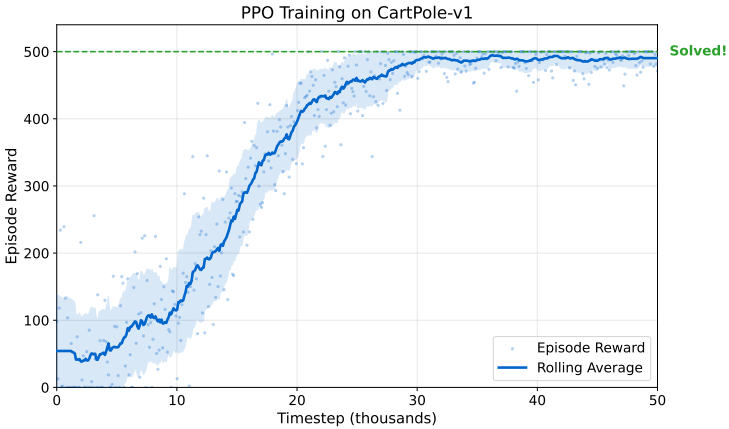
The Modern RL Stack

Layer	Tool
Environments	Gymnasium (successor to OpenAI Gym)
Algorithms	Stable-Baselines3 (PPO, SAC, DQN pre-built)
Training	GPU-accelerated (thousands of episodes/second)
Monitoring	Weights & Biases, TensorBoard
Deployment	ONNX export, sim-to-real transfer

Everything you need is open-source and pip-installable.

pip install gymnasium stable-baselines3. That is all you need to start training RL agents today.

PPO: The Workhorse Algorithm



- PPO learns fast: solves CartPole in ~30k timesteps
- Stable updates: no catastrophic performance drops
- The default choice for most RL tasks in 2025

PPO = Proximal Policy Optimization (Schulman et al., 2017). Used by OpenAI, DeepMind, and most practitioners.

RLHF — How ChatGPT Learned to Be Helpful

1. **Train a base LLM** on text data (supervised learning)
2. **Humans rank outputs** — which response is better?
3. **RL (PPO!) fine-tunes** the LLM to match human preferences

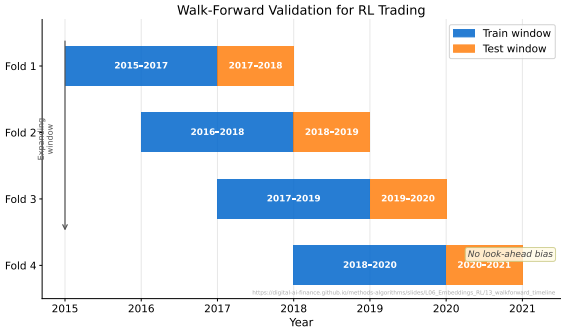
RLHF = Reinforcement Learning from Human Feedback

The reward signal comes from human preferences, not a score function.

The same PPO from Slide 7 is used to align ChatGPT.

InstructGPT paper (Ouyang et al., 2022) introduced RLHF for language models. Now standard for all frontier LLMs.

Where Modern RL Shines



- **Game AI**: AlphaGo, OpenAI Five
- **Robotics**: sim-to-real transfer
- **Finance**: walk-forward validated trading

Train millions of episodes in simulation (free, fast, safe), then deploy.

Sim-to-real: train in simulation where mistakes are free, then fine-tune on real data.

Try It Yourself — Five Lines of Python

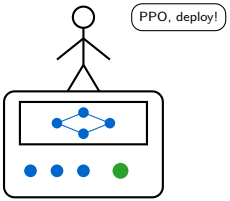
```
from stable_baselines3 import PPO
import gymnasium as gym
env = gym.make("CartPole-v1")
model = PPO("MlpPolicy", env)
model.learn(total_timesteps=50_000)
```

Five lines. That is a complete RL training pipeline.

---

pip install stable-baselines3[extra]. CartPole is the "Hello World" of RL. Try LunarLander-v3 next!

Key Takeaways



1. Modern RL uses neural networks, not tables (Deep RL)
2. PPO is the workhorse algorithm; RLHF uses it to align LLMs
3. Gymnasium + Stable-Baselines3 = 5 lines to a trained agent

XKCD #1838 by Randall Munroe (CC BY-NC 2.5). Next: try the RL notebook to train your own PPO agent!