



Reinforcement Learning

Lecture 23: Model-Based, Max-Entropy

11 Nov 2025

Instructor: Raunak P. Bhattacharyya



ScAI

YARDI SCHOOL OF ARTIFICIAL INTELLIGENCE
INDIAN INSTITUTE OF TECHNOLOGY DELHI

Plan

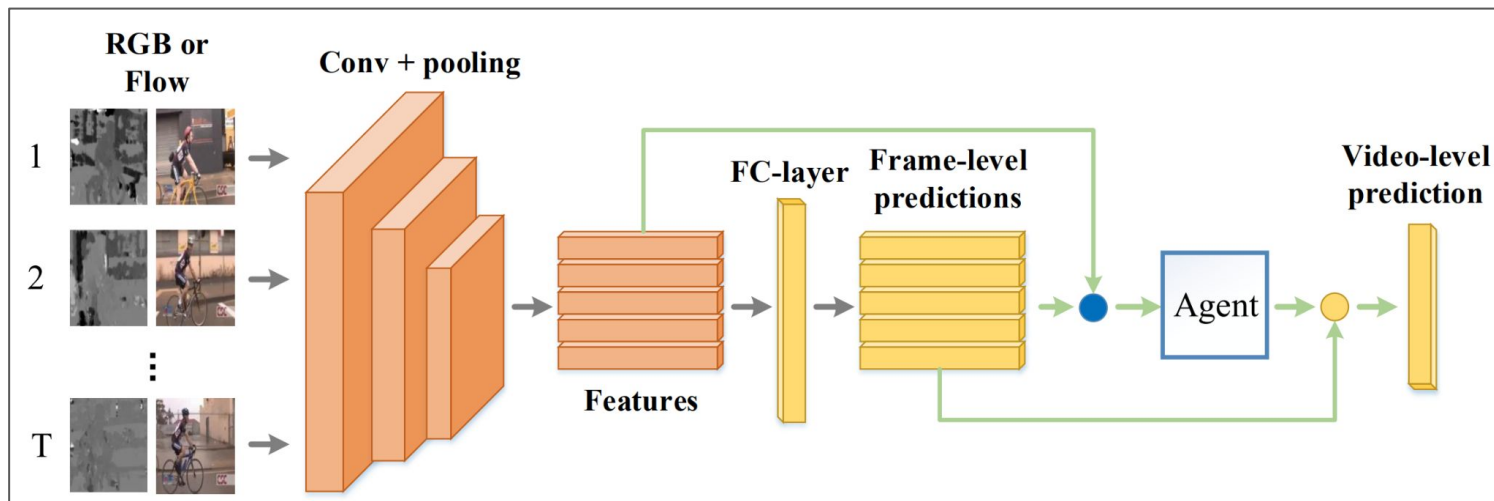
- PO applications
- Model-based RL
- Max-Ent Framework

Policy Optimisation Applications

Reinforce: Application

Attention-Aware Sampling via Deep Reinforcement Learning for Action Recognition

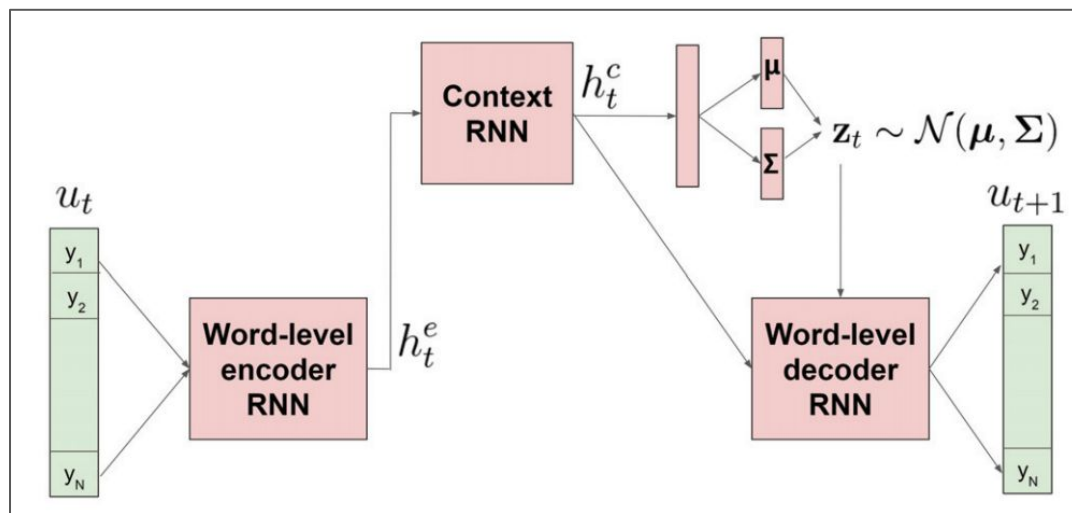
Wenkai Dong,^{1,3} Zhaoxiang Zhang,^{1,2,3*} Tieniu Tan^{1,2,3}



Reinforce: Application

Hierarchical Reinforcement Learning for Open-Domain Dialog

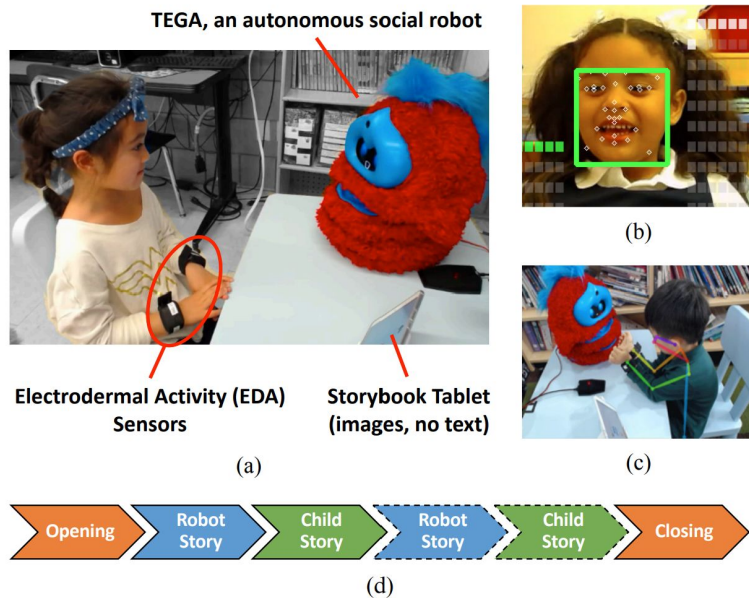
Abdelrhman Saleh,^{1,*} Natasha Jaques,^{2,*} Asma Ghandeharioun,²
Judy Hanwen Shen,² Rosalind Picard²



Reinforce: Application

A Model-Free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education

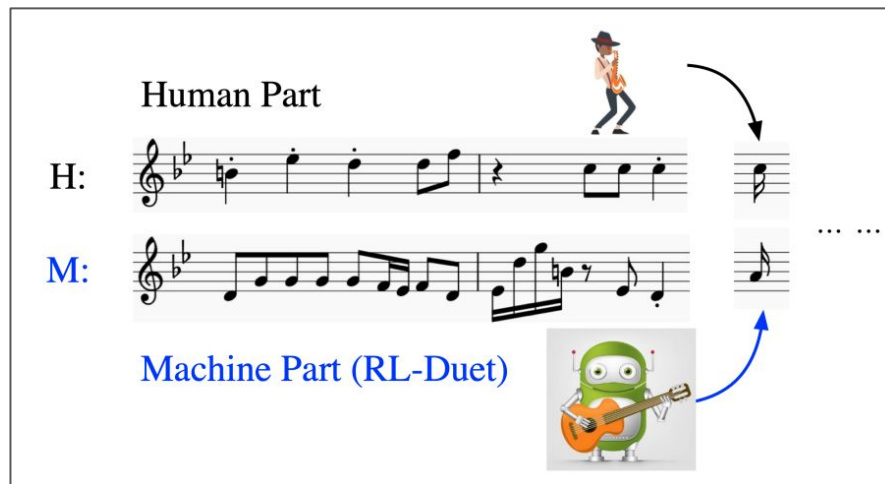
Hae Won Park,¹ Ishaan Grover,¹ Samuel Spaulding,¹ Louis Gomez,^{1,2} Cynthia Breazeal¹



PPO: Application

RL-Duet: Online Music Accompaniment Generation Using Deep Reinforcement Learning

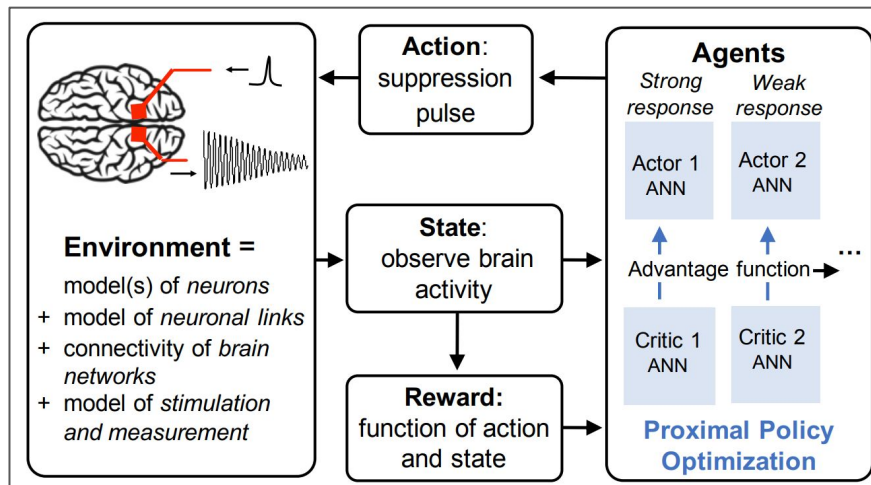
Nan Jiang¹ Sheng Jin¹ Zhiyao Duan² Changshui Zhang¹



PPO: Application

Reinforcement Learning Framework for Deep Brain Stimulation Study

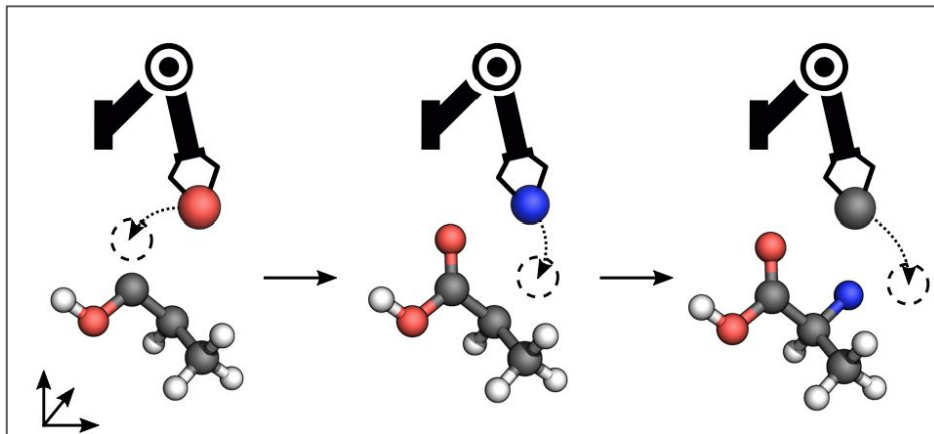
Dmitrii Krylov¹, Remi Tachet des Combes², Romain Laroche²,
Michael Rosenblum³ and Dmitry V. Dylov¹



PPO: Application

Reinforcement Learning for Molecular Design Guided by Quantum Mechanics

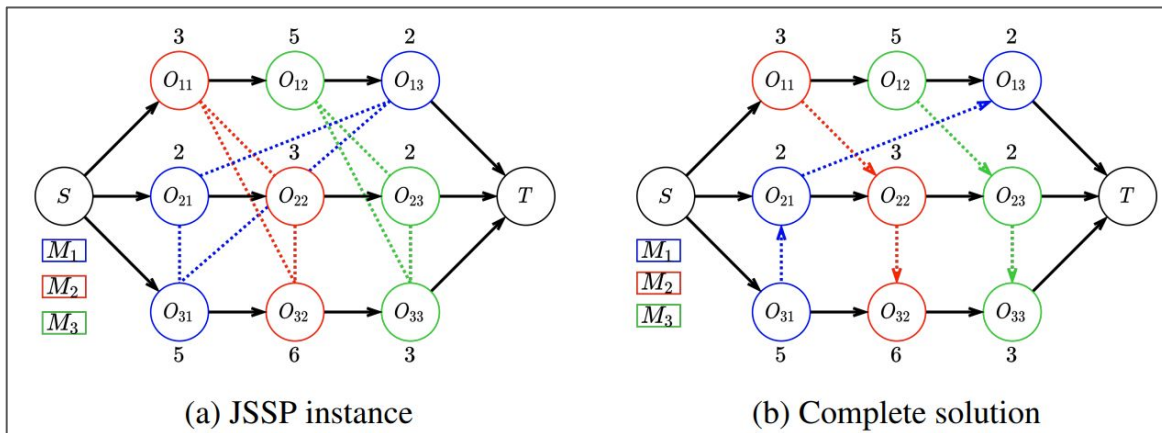
Gregor N. C. Simm ^{*1} Robert Pinsler ^{*1} José Miguel Hernández-Lobato ¹



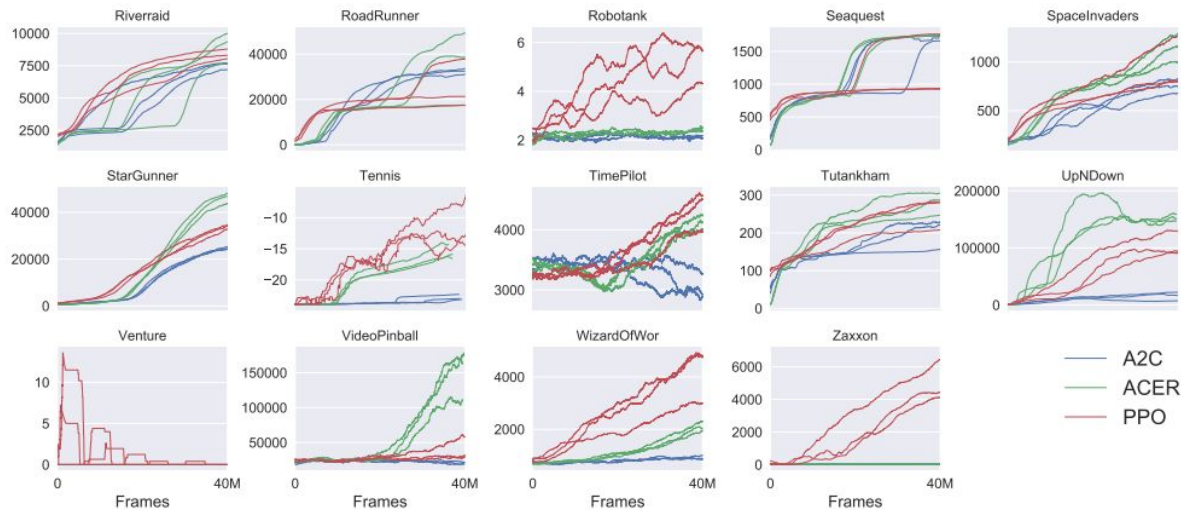
PPO: Application

Learning to Dispatch for Job Shop Scheduling via Deep Reinforcement Learning

Cong Zhang^{1,*}, Wen Song^{2,*}, Zhiguang Cao^{3,†}, Jie Zhang¹, Puay Siew Tan⁴, and Chi Xu⁴

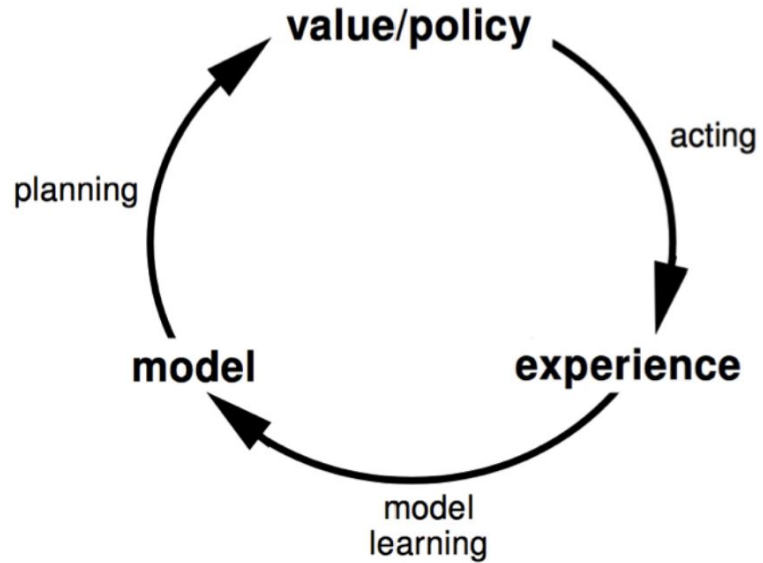


Policy Opt is Not Enough

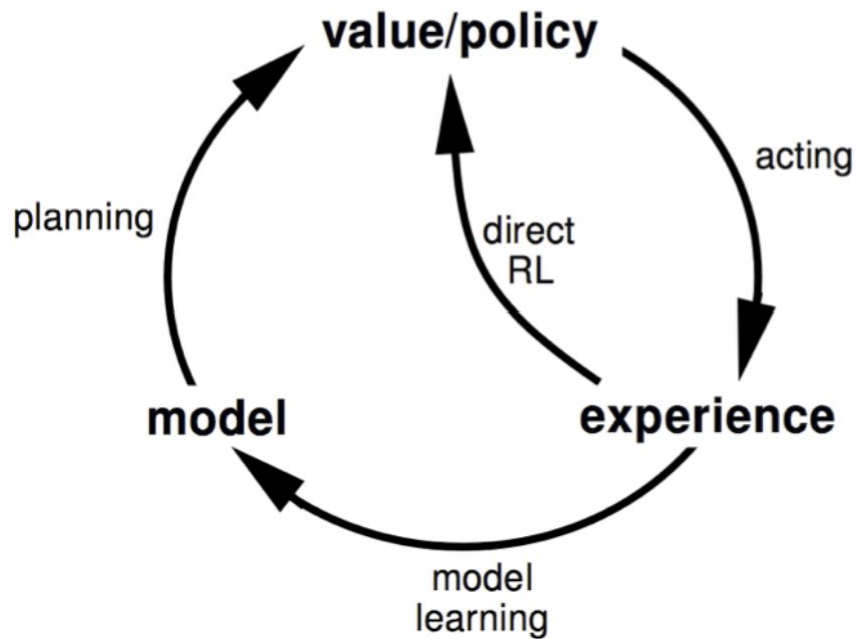


Model-Based RL

Key Idea



Motivation



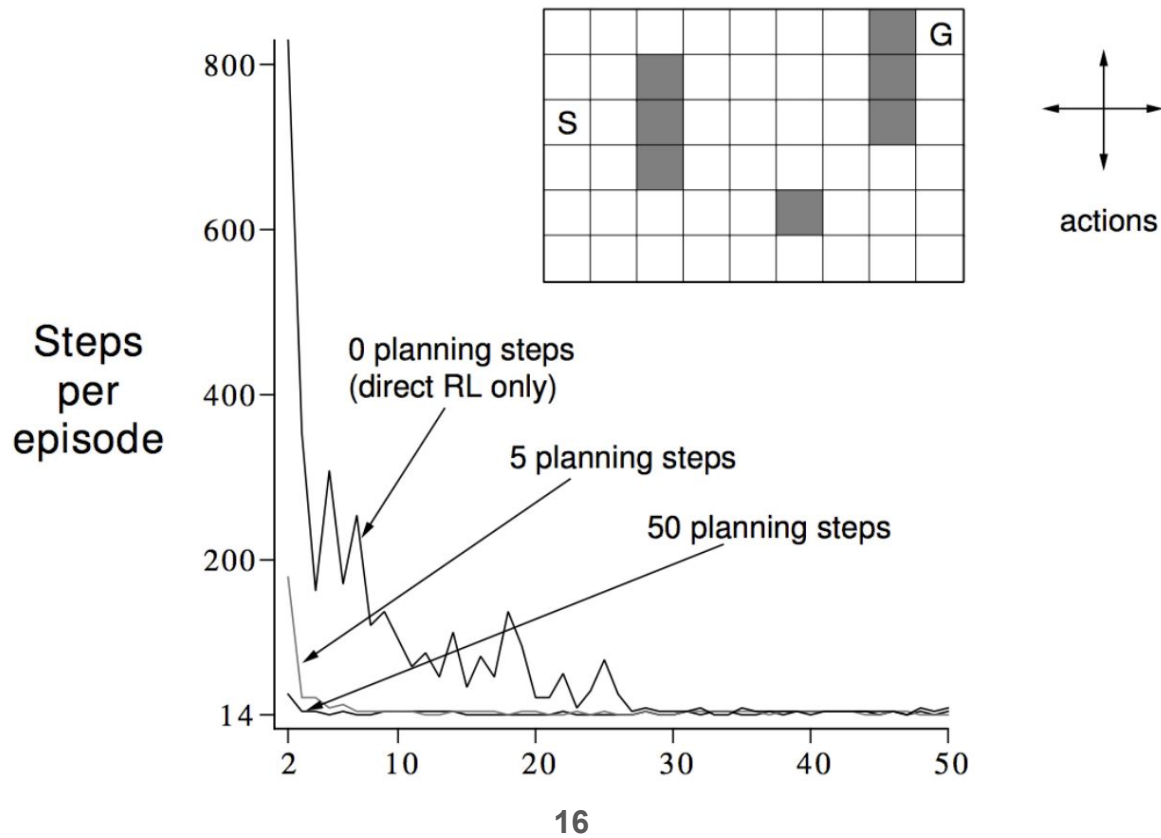
Integrated Learning and Planning

Initialize $Q(s, a)$ and $Model(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$

Do forever:

- (a) $S \leftarrow$ current (nonterminal) state
- (b) $A \leftarrow \varepsilon$ -greedy(S, Q)
- (c) Execute action A ; observe resultant reward, R , and state, S'
- (d) $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$
- (e) $Model(S, A) \leftarrow R, S'$ (assuming deterministic environment)
- (f) Repeat n times:
 - $S \leftarrow$ random previously observed state
 - $A \leftarrow$ random action previously taken in S
 - $R, S' \leftarrow Model(S, A)$
 - $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$

Dyna-Q on Maze Task



Max Entropy Formulation