0

APRENDIZAJE POR REFUERZO I (CURSADO)

DOCENTE DEL CURSO

- Miguel Augusto Azar
- Ingeniero en Informática
- Especialista en Docencia Superior
- Docente e investigador

¿EN QUÉ CONSISTE APRENDIZAJE POR REFUERZO I?

- Consiste en el desarrollo de algoritmos que aprenden por sí mismos (o asistidos) una determinada tarea.

CASOS DE USO

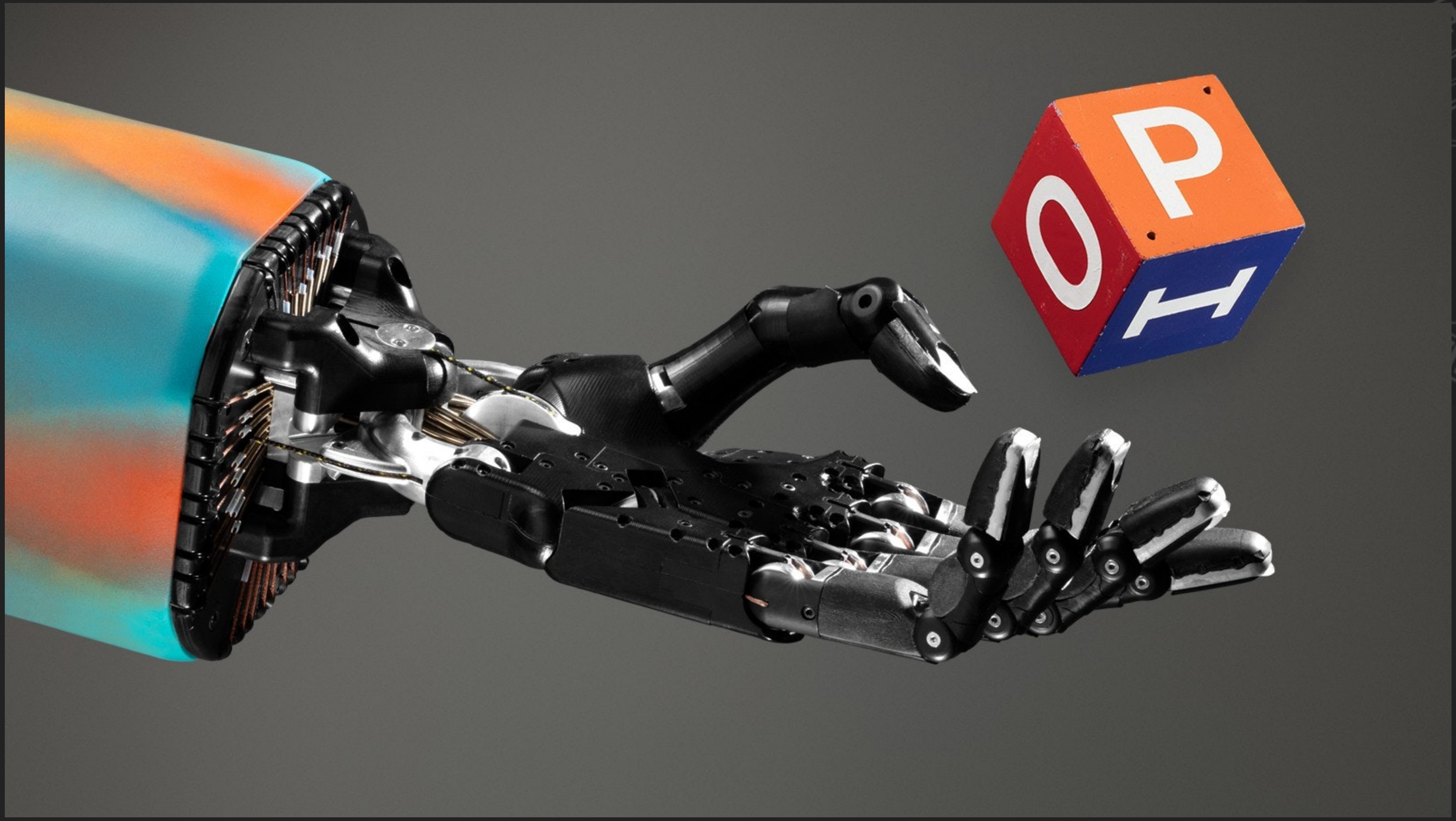
- Vehículo autónomo que aprende a conducir por sí mismo.

CASOS DE USO

- Learning to drive in a day (2018)
- DDPG (Deep Deterministic Policy Gradient)
- https://www.youtube.com/watch?v=eRwTbRtnT1I

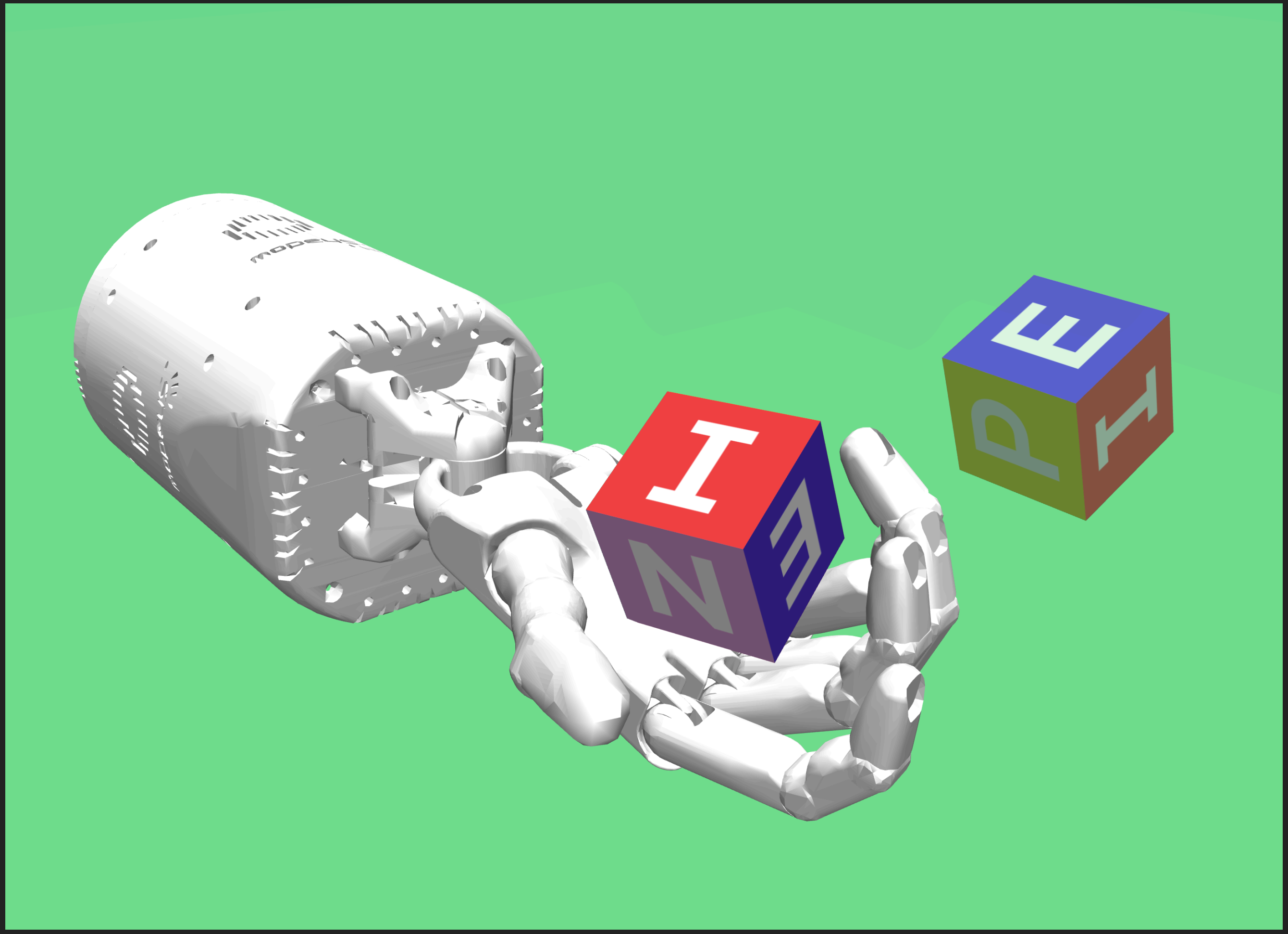
CASOS DE USO

- Mano robótica que aprende a manipular objetos.



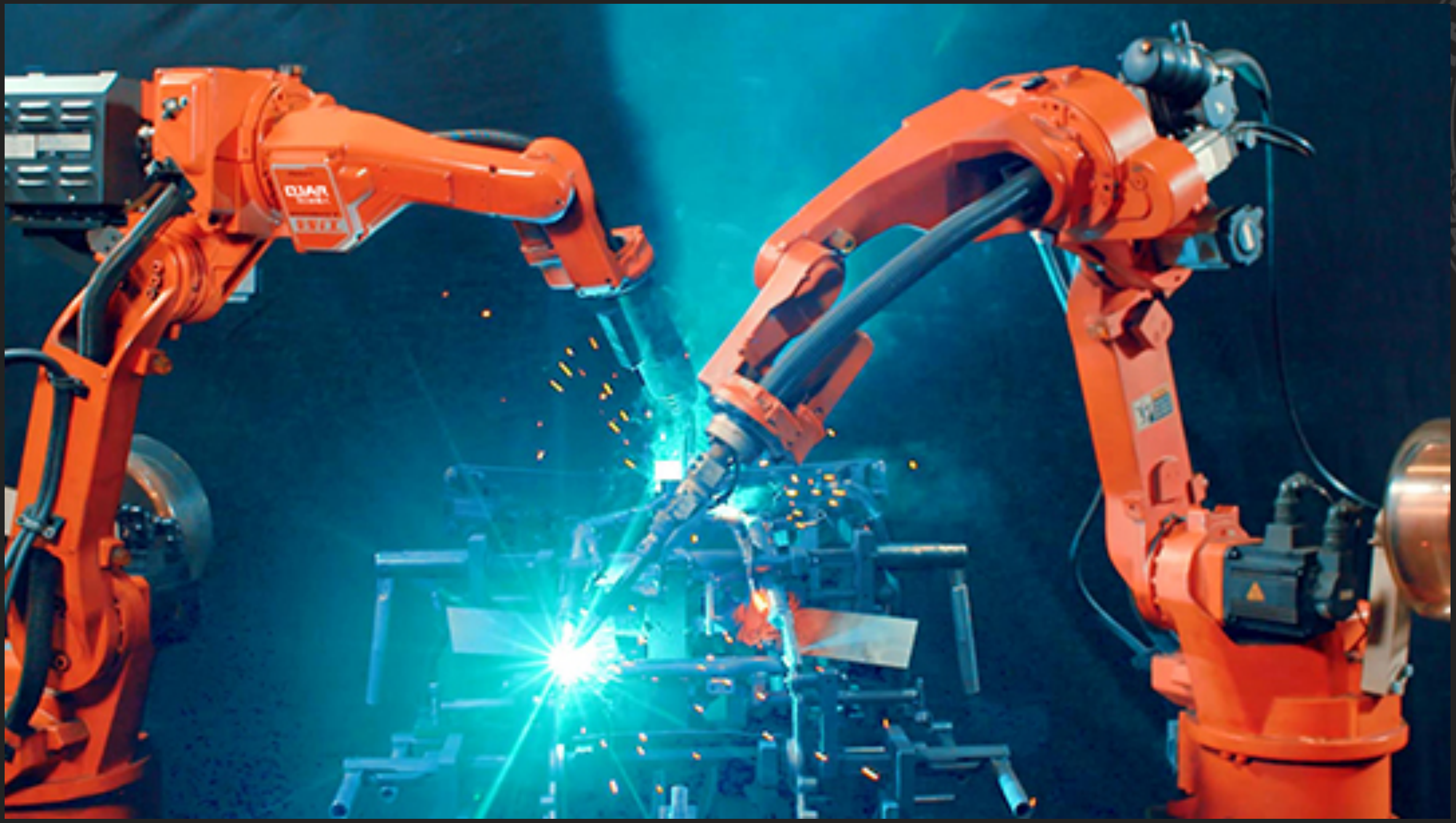
CASOS DE USO

- Learning Dexterity (OpenAI) (2018)
- https://www.youtube.com/watch?v=jwSbzNHGflM



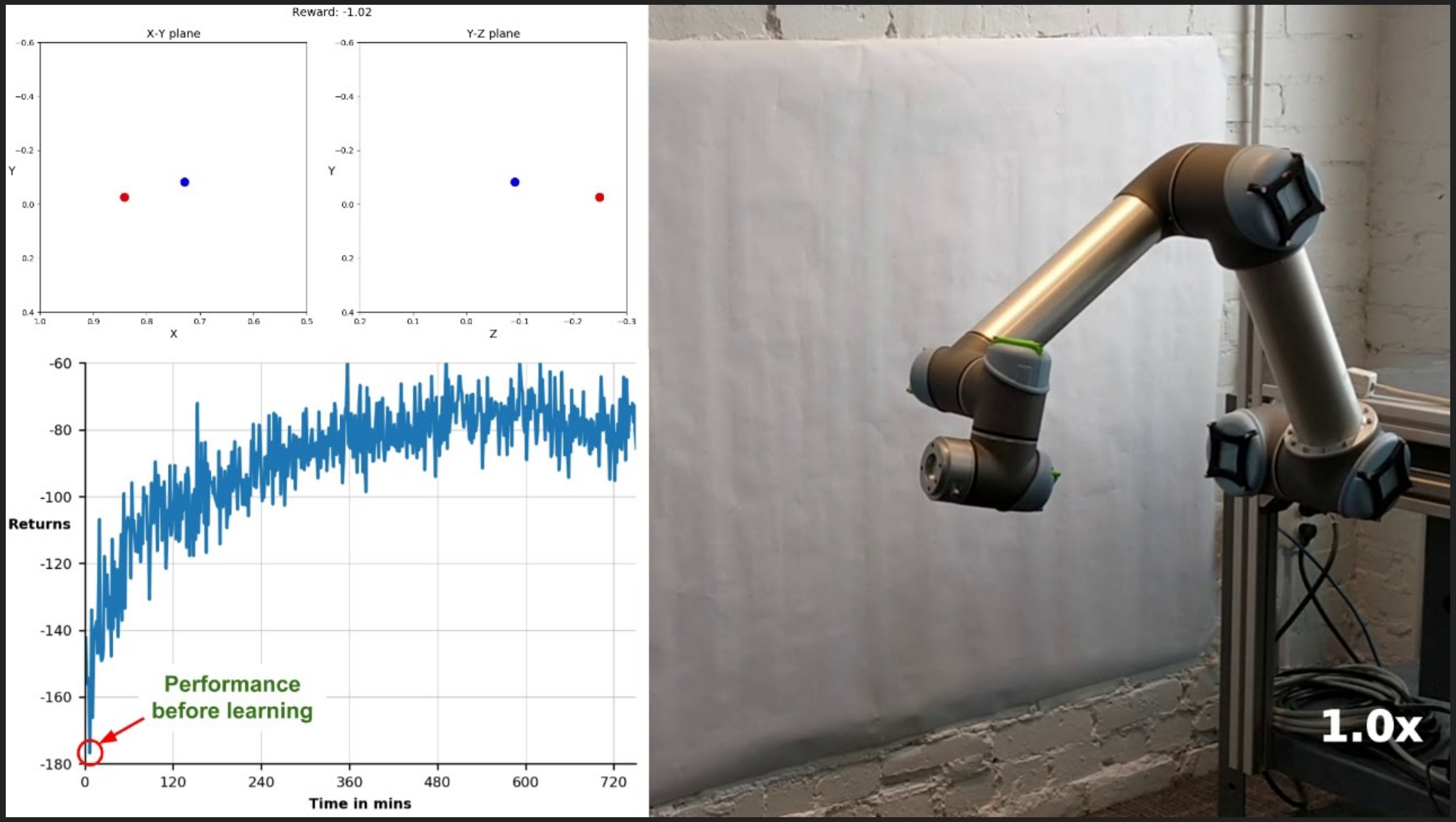
CASOS DE USO

- Brazo robótico que aprende a desplazarse hacia un punto en el espacio..



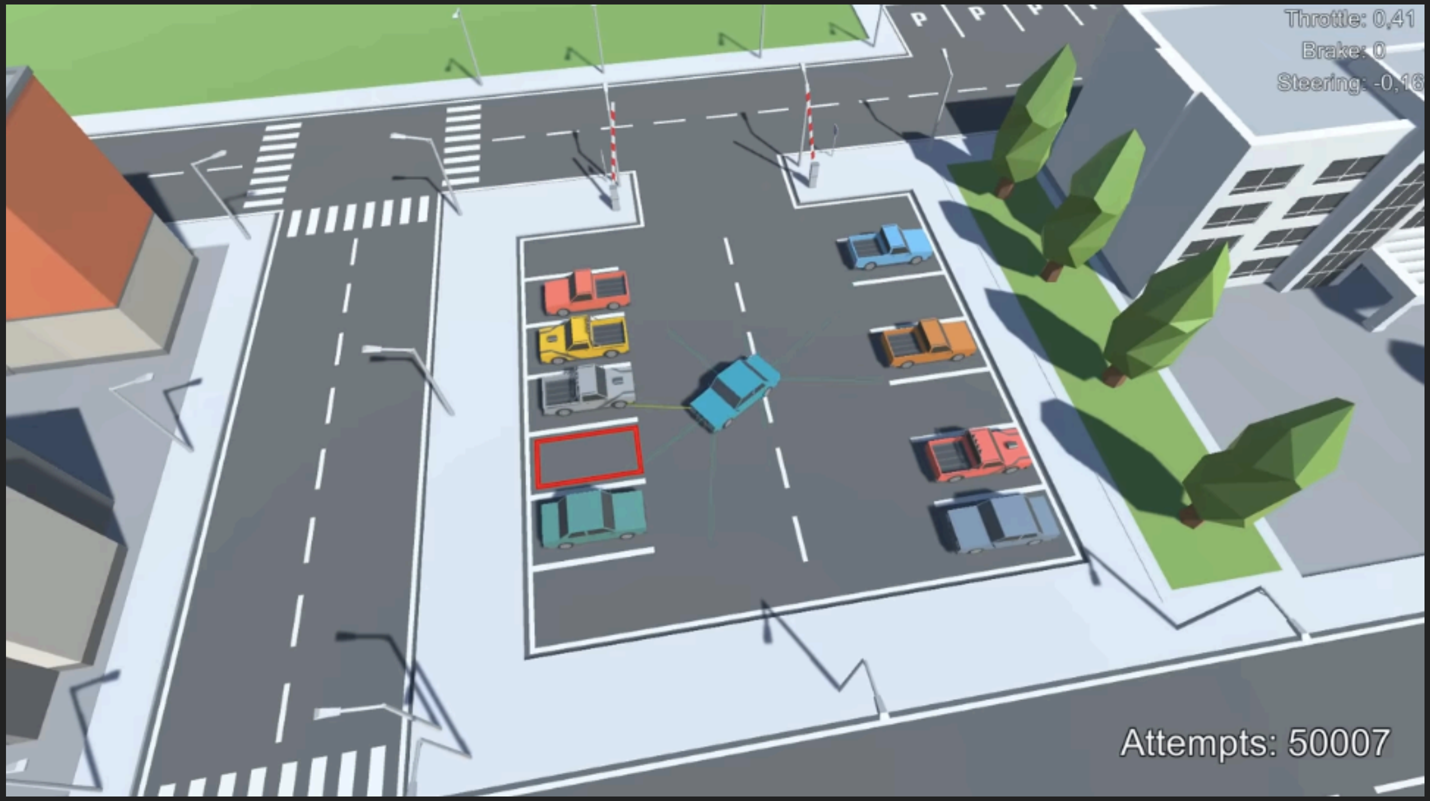
CASOS DE USO

- Setting up a Reinforcement Learning Task with a Real-World Robot. (Mahmood et al., 2018)
- https://www.youtube.com/watch?v=ZVIxt2rt1_4



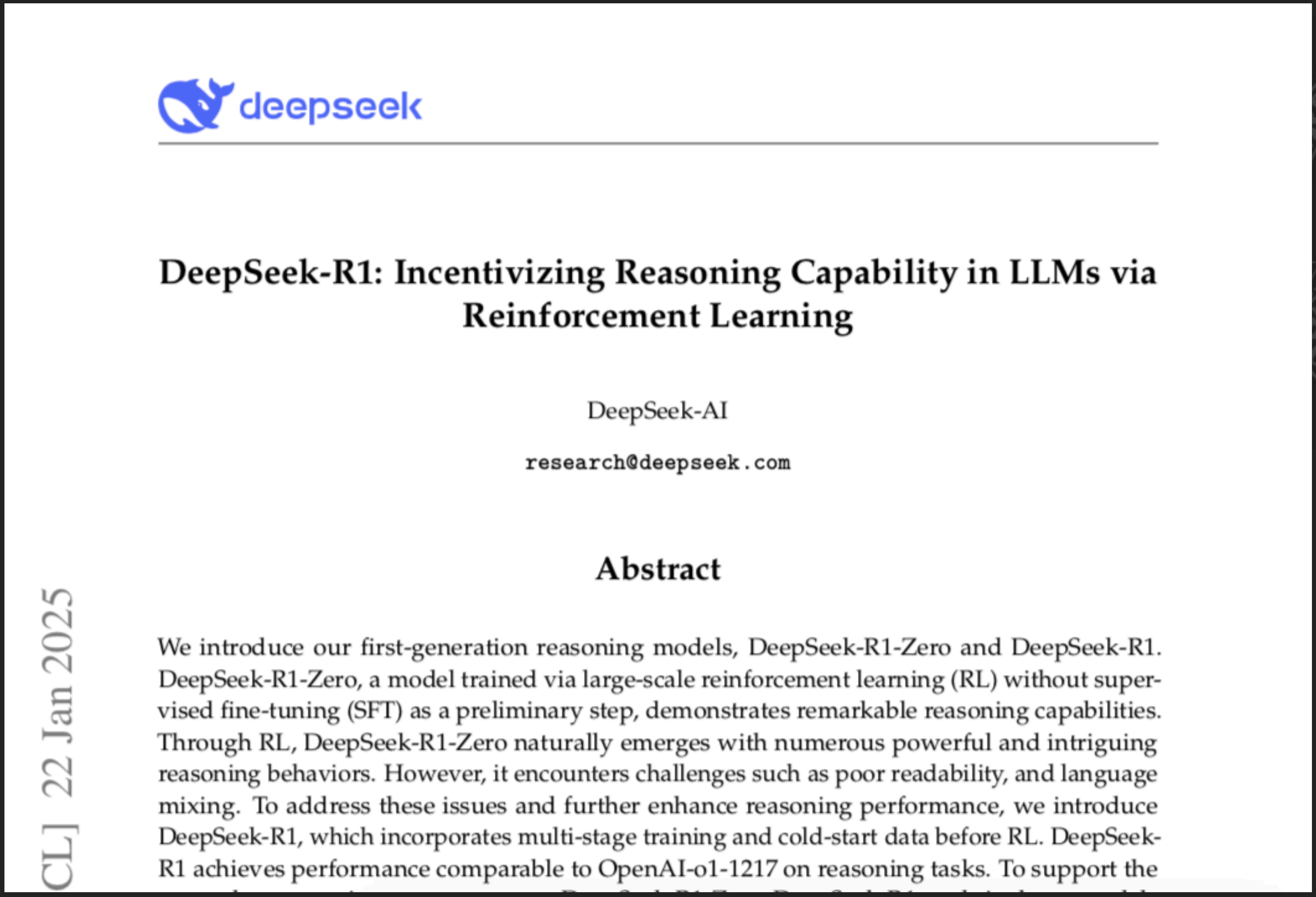
CASOS DE USO

- AI Learns to Park (2019)
- PPO (Proximal Policy Optimization)
- https://www.youtube.com/watch?v=VMp6pq6_QjI



CASOS DE USO

- Deepseek-R1 (2025)


deepseek

CL] 22 Jan 2025

DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning

DeepSeek-AI

research@deepseek.com

Abstract

We introduce our first-generation reasoning models, DeepSeek-R1-Zero and DeepSeek-R1. DeepSeek-R1-Zero, a model trained via large-scale reinforcement learning (RL) without supervised fine-tuning (SFT) as a preliminary step, demonstrates remarkable reasoning capabilities. Through RL, DeepSeek-R1-Zero naturally emerges with numerous powerful and intriguing reasoning behaviors. However, it encounters challenges such as poor readability, and language mixing. To address these issues and further enhance reasoning performance, we introduce DeepSeek-R1, which incorporates multi-stage training and cold-start data before RL. DeepSeek-R1 achieves performance comparable to OpenAI-o1-1217 on reasoning tasks. To support the

CASOS DE USO

- Deepseek-R1 (2025)

2.2.1. *Reinforcement Learning Algorithm*

Group Relative Policy Optimization In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \quad (1)$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \quad (2)$$

where ε and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \quad (3)$$

5

CASOS DE USO

- LLMs mejorados con Aprendizaje por Refuerzo [Wang et al., 2024]

RL Enhanced LLMs	Organization	# Params	RL Methods
Instruct-GPT (Ouyang et al., 2022)	OpenAI	1.3B, 6B, 175B	RLHF, PPO
GPT-4 (OpenAI, 2023)	OpenAI	-	RLHF, PPO, RBRM
Gemini (Team et al., 2023)	Google	-	RLHF
InternLM2 (Cai et al., 2024)	上海人工智能实验室	1.8B, 7B, 20B	RLHF, PPO
Claude 3 (Anthropic, 2024)	ANTHROPIC	-	RLAIF
Reka (Team et al., 2024c)	Reka	7B, 21B	RLHF, PPO
Zephyr (HuggingFaceH4, 2024)	Argilla	141B-A39B	ORPO
Phi-3 (Abdin et al., 2024)	Microsoft	3.8B, 7B, 14B	DPO
DeepSeek-V2 (Liu et al., 2024a)	deepseek	236B-A21B	GRPO
ChatGLM (GLM et al., 2024)	ZHIPU·AI	6B, 9B	ChatGLM-RLHF
Nemotron-4 340B (Adler et al., 2024)	NVIDIA	340B	DPO, RPO
Llama 3 (Dubey et al., 2024)	Meta	8B, 70B, 405B	DPO
Qwen2 (Yang et al., 2024a)	Alibaba	(0.5-72)B, 57B-A14B	DPO
Gemma2 (Team et al., 2024b)	Google	2B, 9B, 27B	RLHF
Starling-7B (Zhu et al., 2024)	Berkeley	7B	RLAIF, PPO
Athene-70B (Nexusflow, 2024)	Nexusflow	70B	RLHF
Hermes 3 (Teknium et al., 2024)	NOUS	8B, 70B, 405B	DPO
o1 (OpenAI, 2024b)	OpenAI	-	RL through CoT

TEMAS Y TÉCNICAS A ABORDAR EN EL CURSO

- Introducción
- Q-Learning
- Deep Q-Learning
- Actor-Critic
- PPO

CURSADO

- Parte 1 (~ 80 min.)
- Break (~ 10-20 min.)
- Parte 2 (~ 80 min.)

EVALUACIÓN

- Se aprueba la materia implementando 3 desafíos prácticos.
    - ✓ Individual.
    - ✓ Grupal (hasta 3 integrantes).
    - ✓ Informe explicativo de al menos 3 carillas (.pdf).
    - ✓ Código fuente transcripto en el informe y también en un repositorio (enlace al repositorio en el informe, .py, .ipynb, otras extensiones asociadas al lenguaje Python solo si es necesario).
    - ✓ Se pueden presentar durante la cursada o al final.
    - ✓ Los desafíos son abiertos. El estudiante elige un caso de uso según la técnica estudiada y elabora el desafío práctico.

REFERENCIAS BIBLIOGRÁFICAS Y WEB (I)

- S. Wang, S. Zhang, J. Zhang, R. Hu, X. Li, T. Zhang, et al., “Reinforcement Learning Enhanced LLMs: A Survey,” arXiv preprint arXiv:2412.10400, 2024.
- J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” arXiv preprint arXiv:1707.06347, 2017.
- Hillier, F. S. Lieberman, G. J. (2010). Introducción a la Investigación de Operaciones.
- D. Guo et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv preprint arXiv:2501.12948, 2025.
- https://www.youtube.com/watch?v=eRwTbRtnT1I (Learning to drive in a day)
- A. Kendall et al., "Learning to drive in a day," 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 2019, pp. 8248-8254, doi: 10.1109/ICRA.2019.8793742. https://ieeexplore.ieee.org/abstract/document/8793742
- https://www.youtube.com/watch?v=VMp6pq6_QjI (AI Learns to Park - Deep Reinforcement Learning)
- Mahmood, A. R., Korenkevych, D., Komer, B. J., & Bergstra, J. (2018, October). Setting up a reinforcement learning task with a real-world robot. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 4635-4640). IEEE.