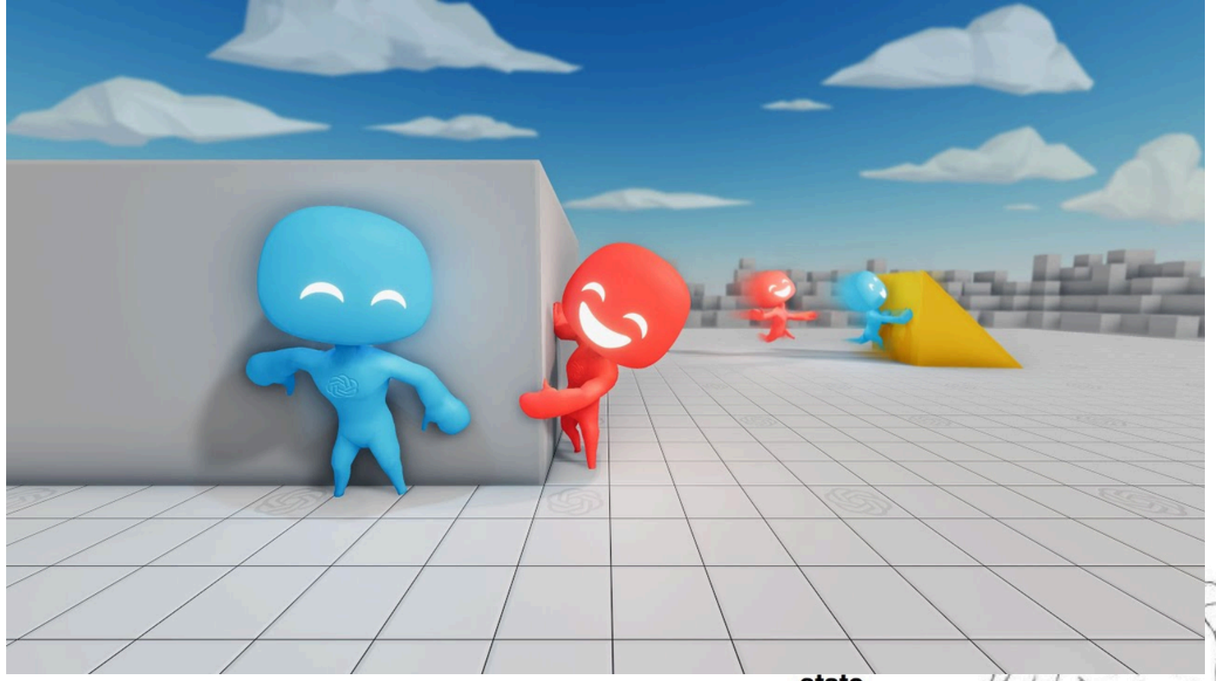
From climbing a mountain to Large Language Models

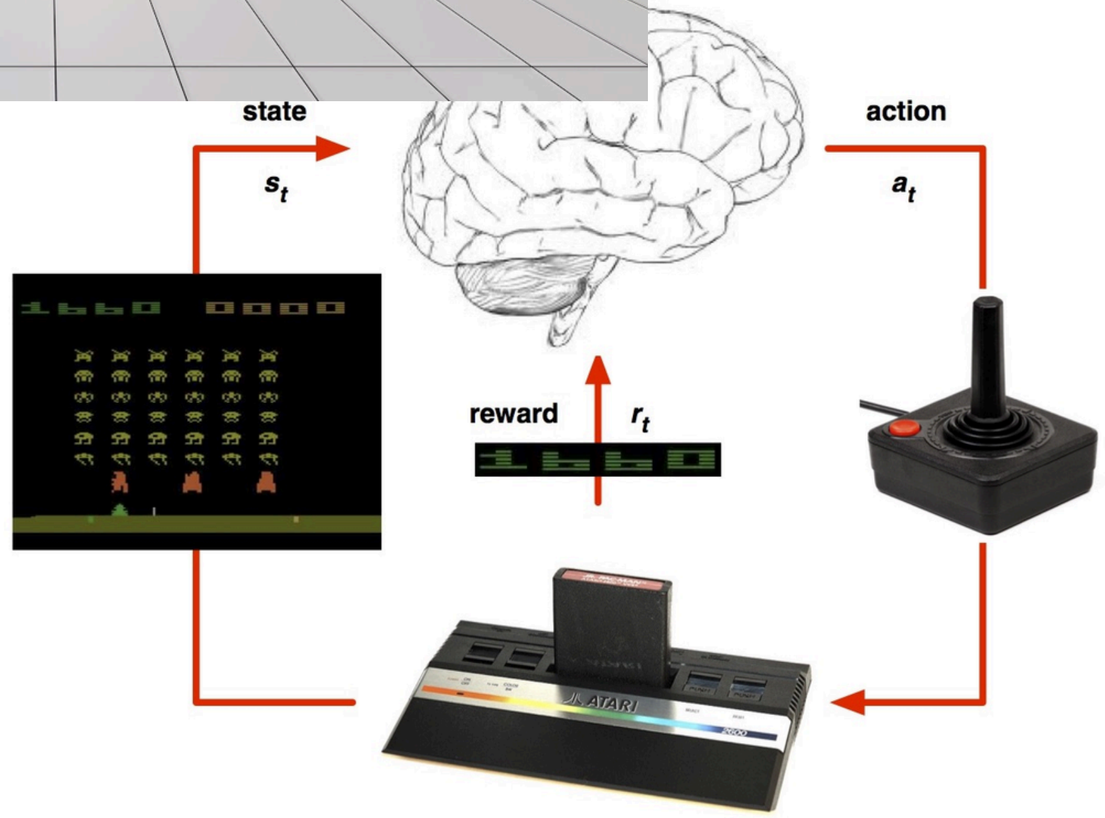
A path of Reinforcement Learning

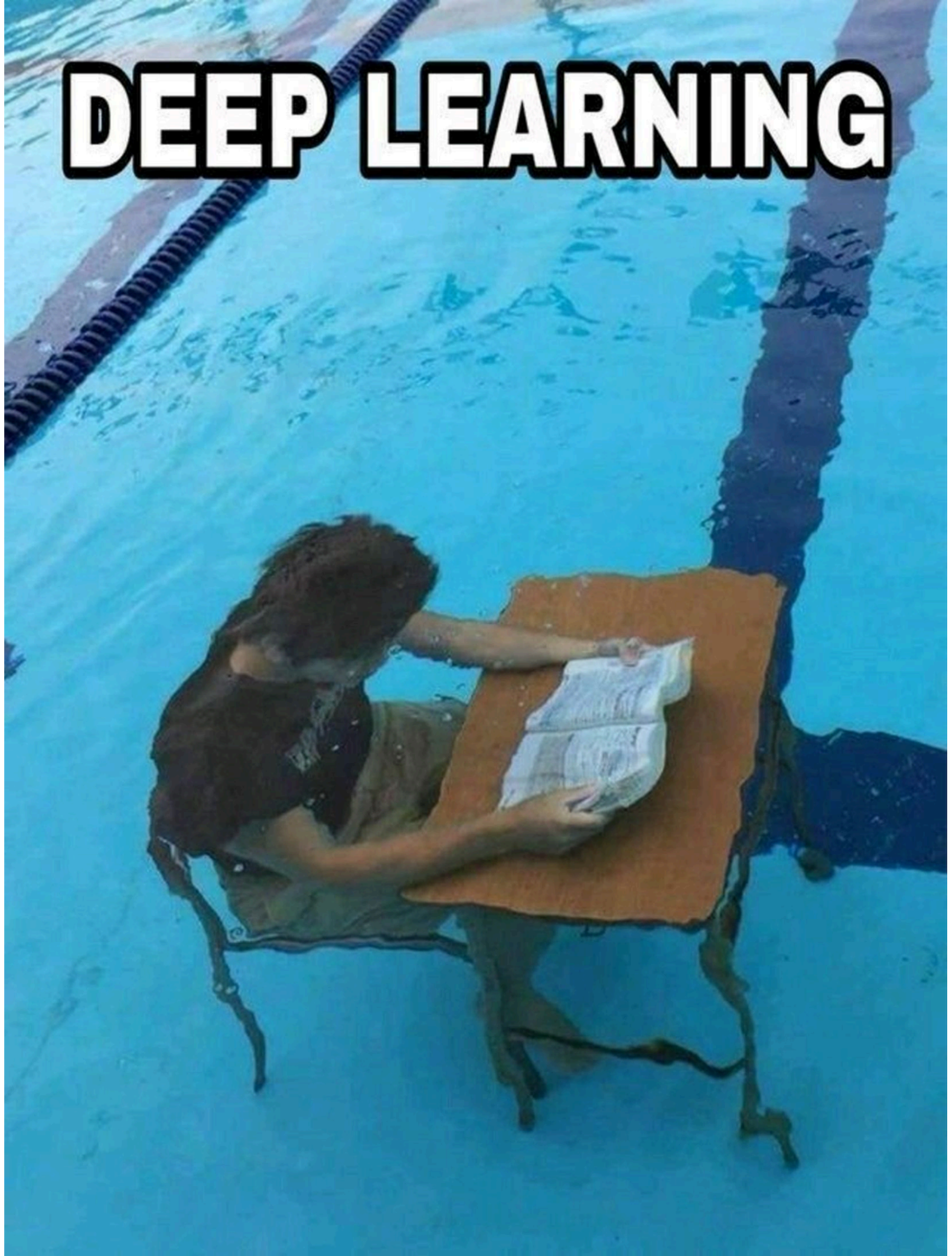
PyData

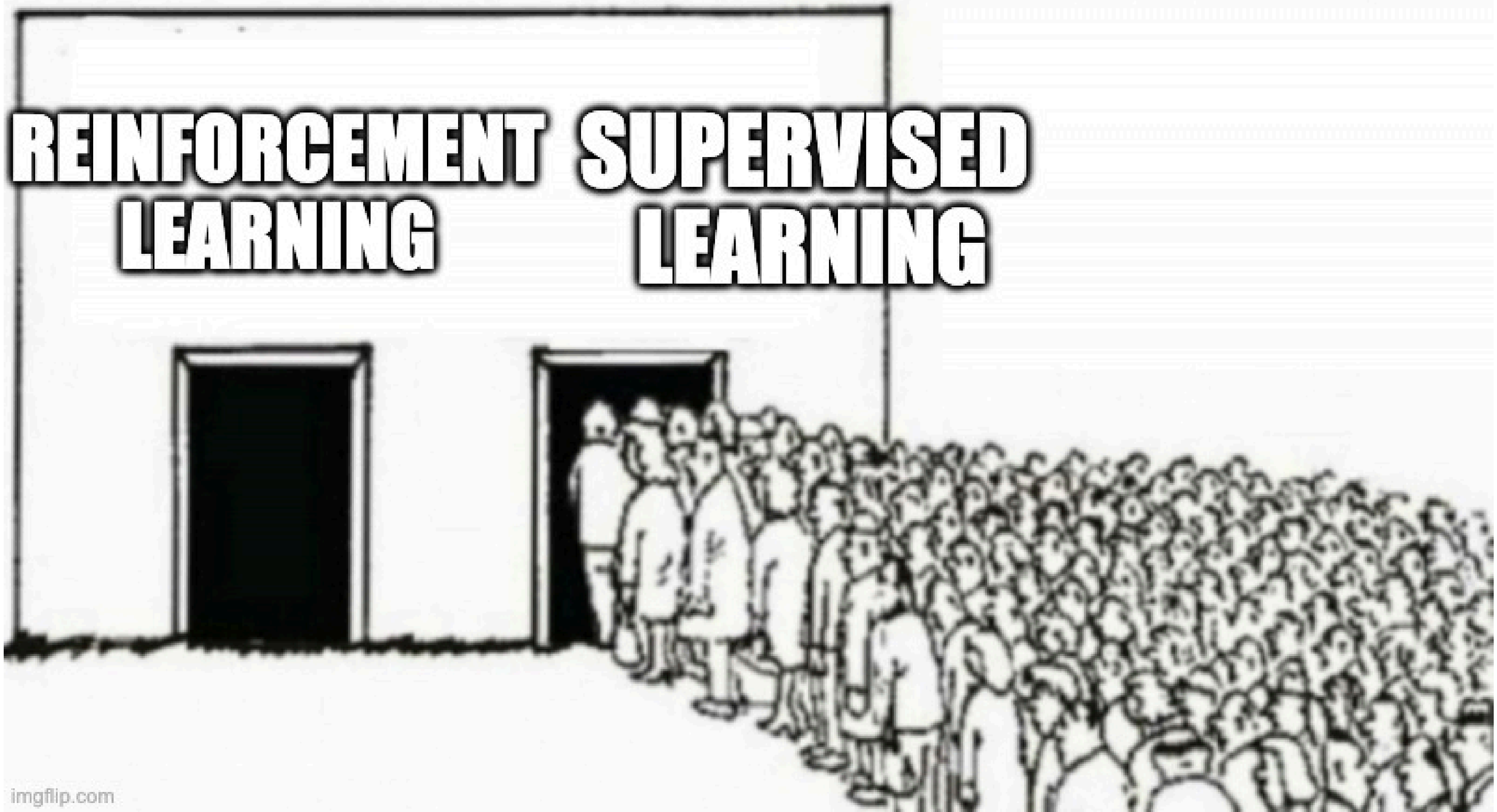
Mali Imre Gergely

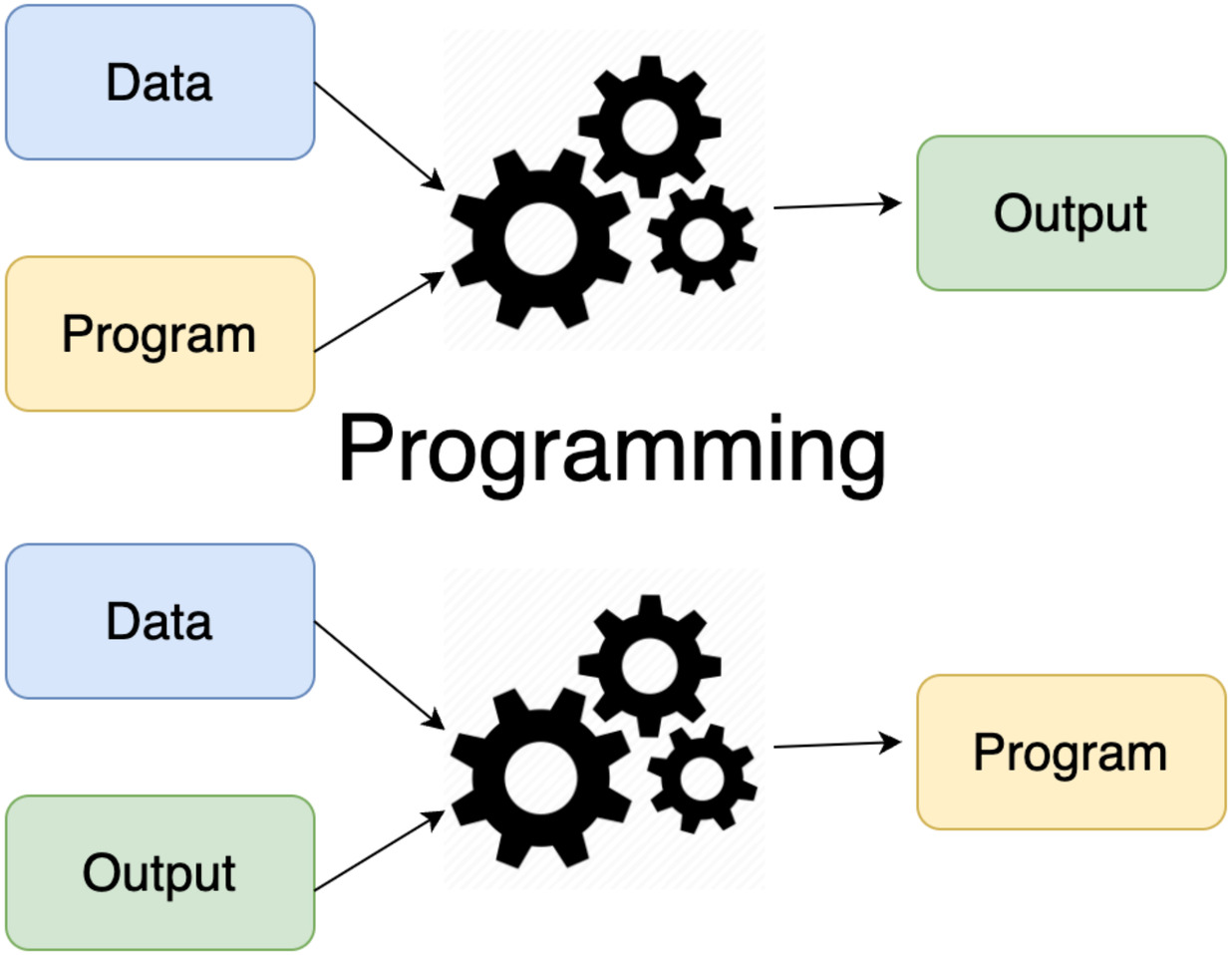


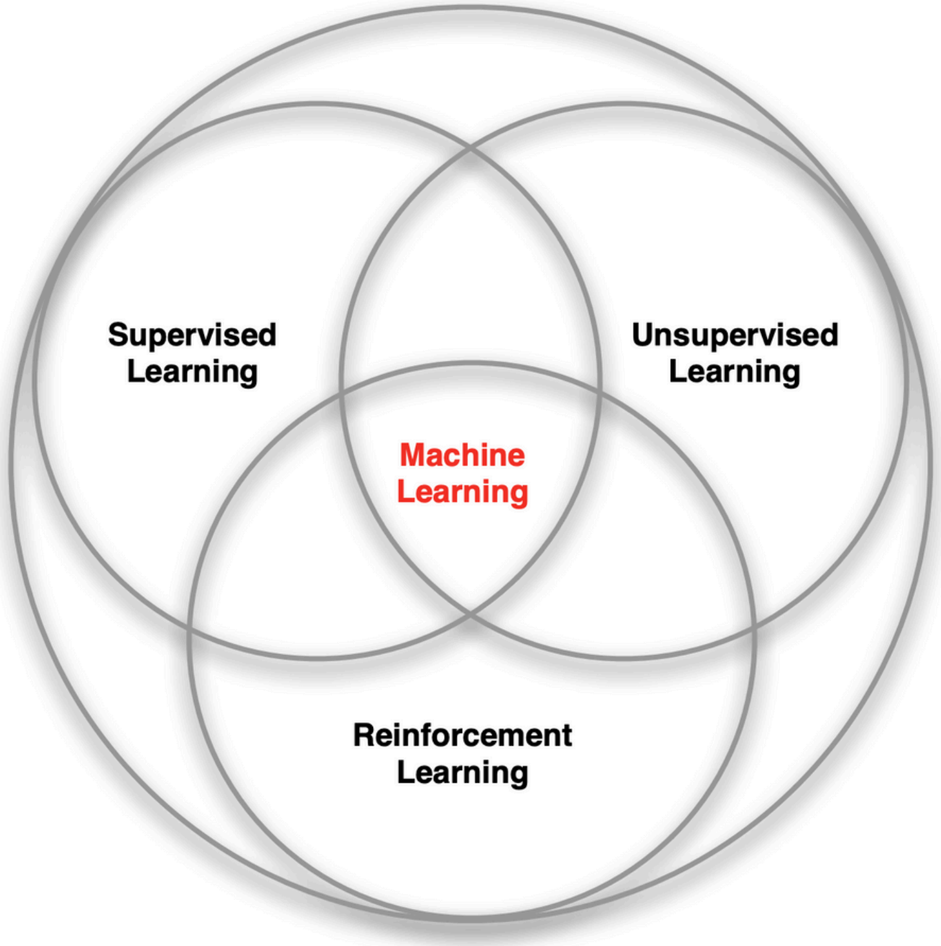

AlphaGo


state
s_t
action
a_t
reward
r_t
ATARI

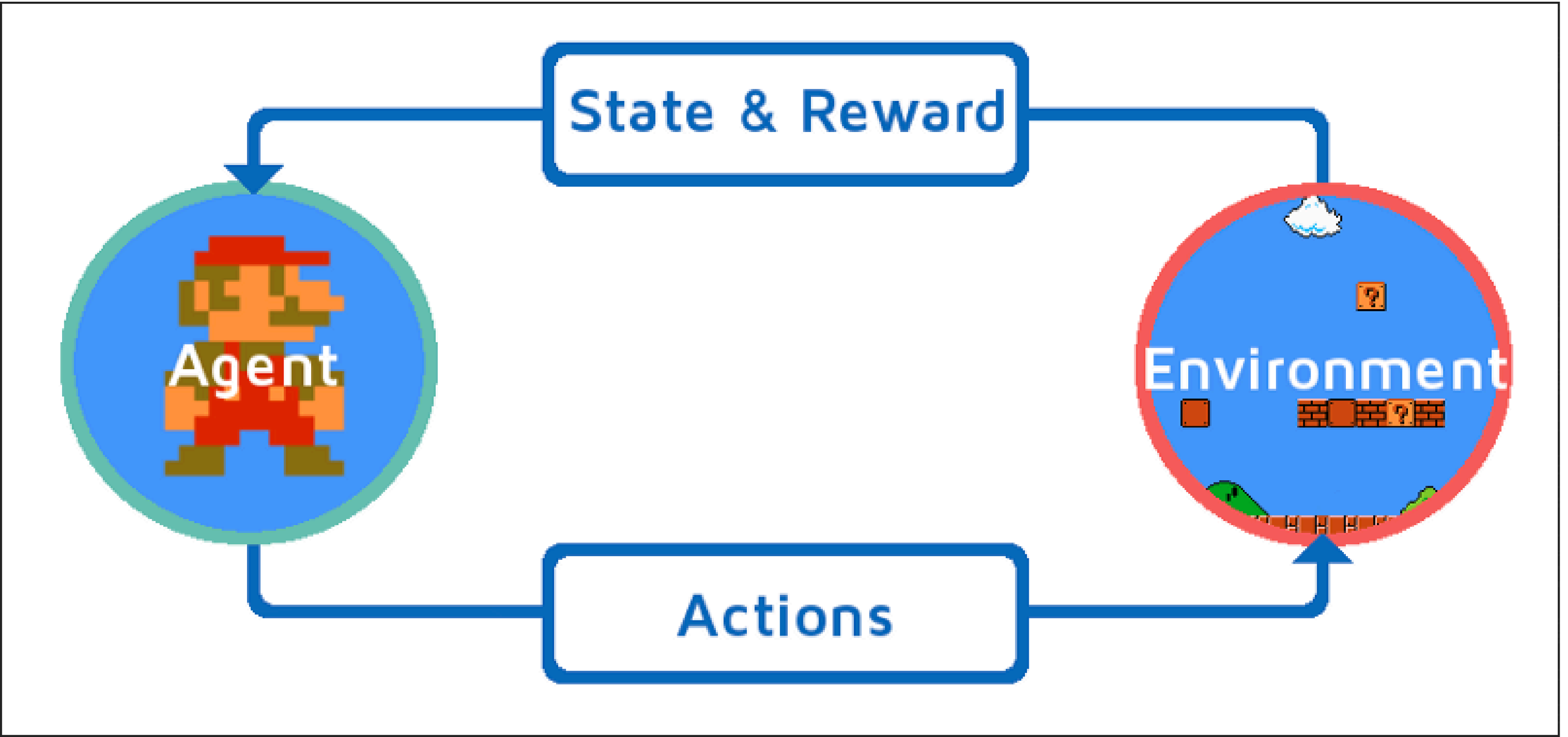

DEEP LEARNING


REINFORCEMENT LEARNING
SUPERVISED LEARNING
imgflip.com


Data
Program
Output
Programming
Data
Output
Program
Machine Learning


Supervised Learning
Unsupervised Learning
Machine Learning
Reinforcement Learning

Agent-Environment Loop



Supervised Learning

Picture1.png - cat

Picture 2.png - dog

Picture3.png - cat

...

RL

Pos(0,0), left, 0.5

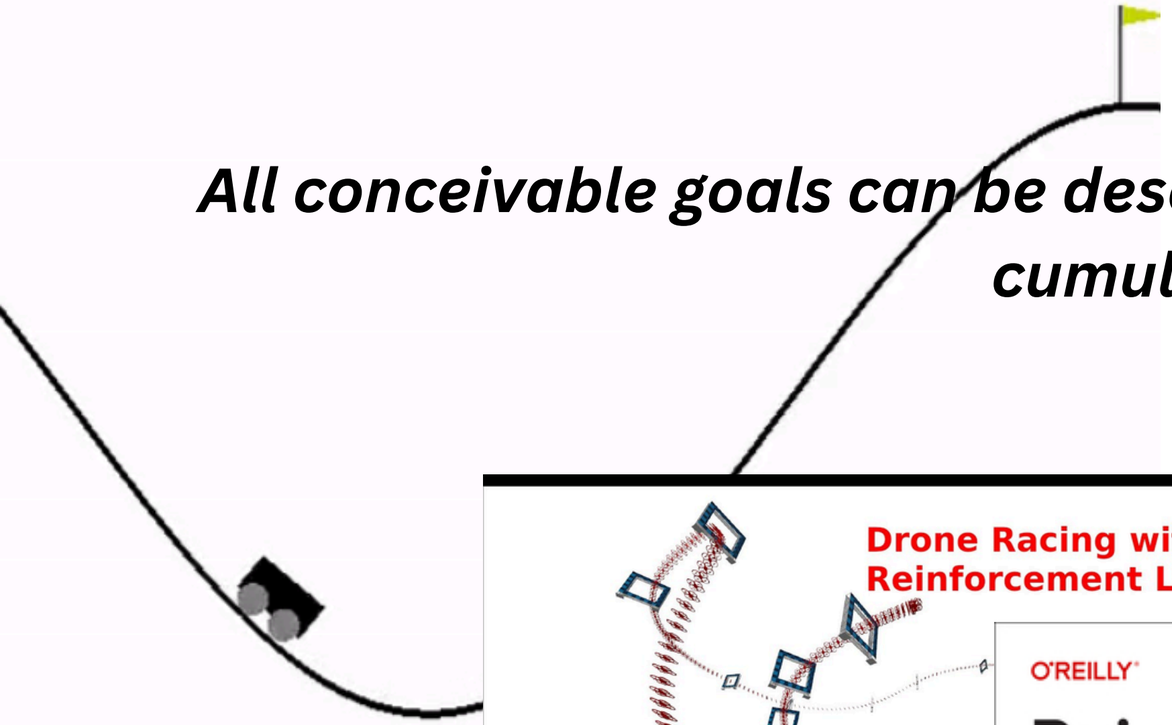
Pos(-1,0), up, -1

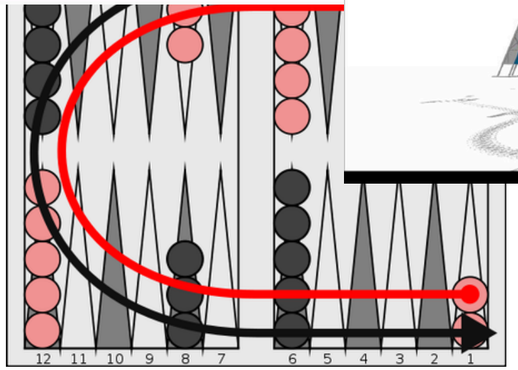
Pos(-1,1), right, 0

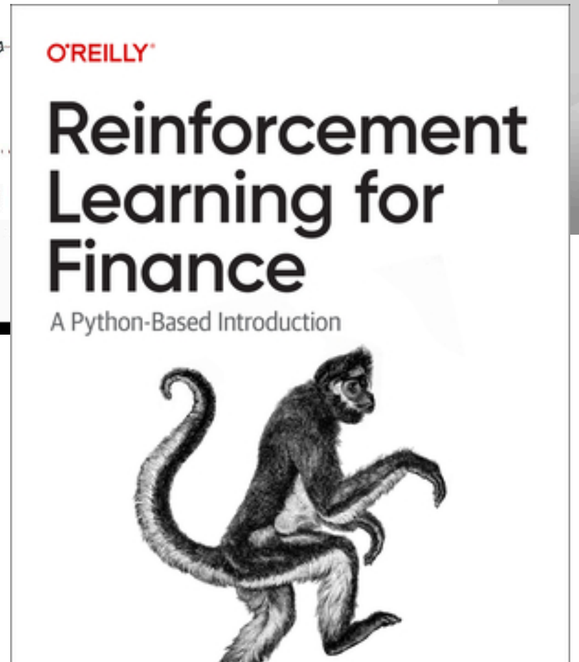
...

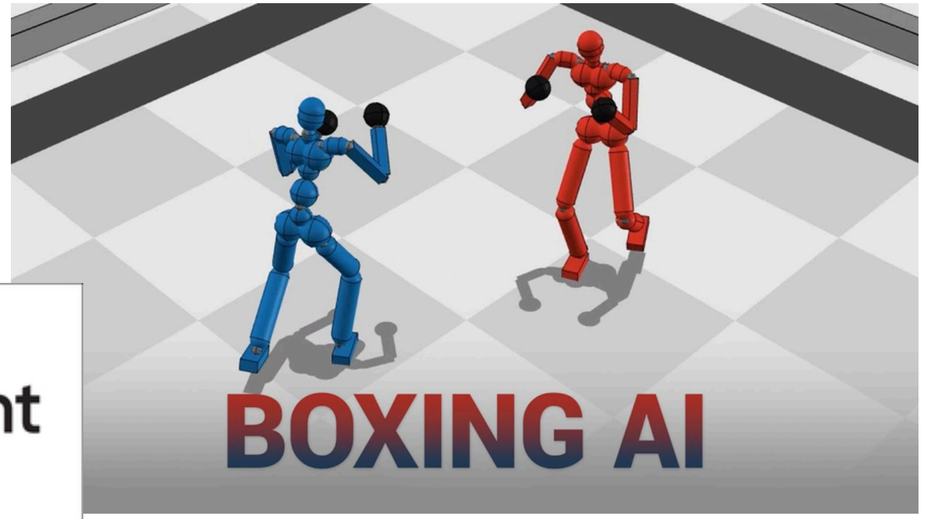
The reward hypothesis

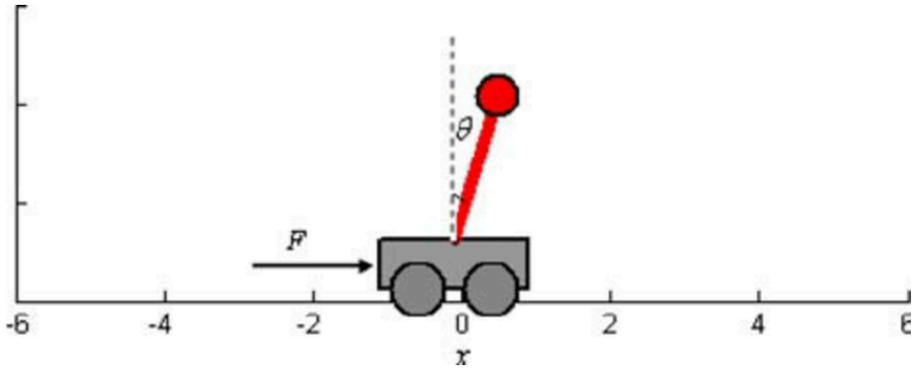
All conceivable goals can be described by the maximisation of expected cumulative reward











Emergence

the whole is more that the sum up of its parts

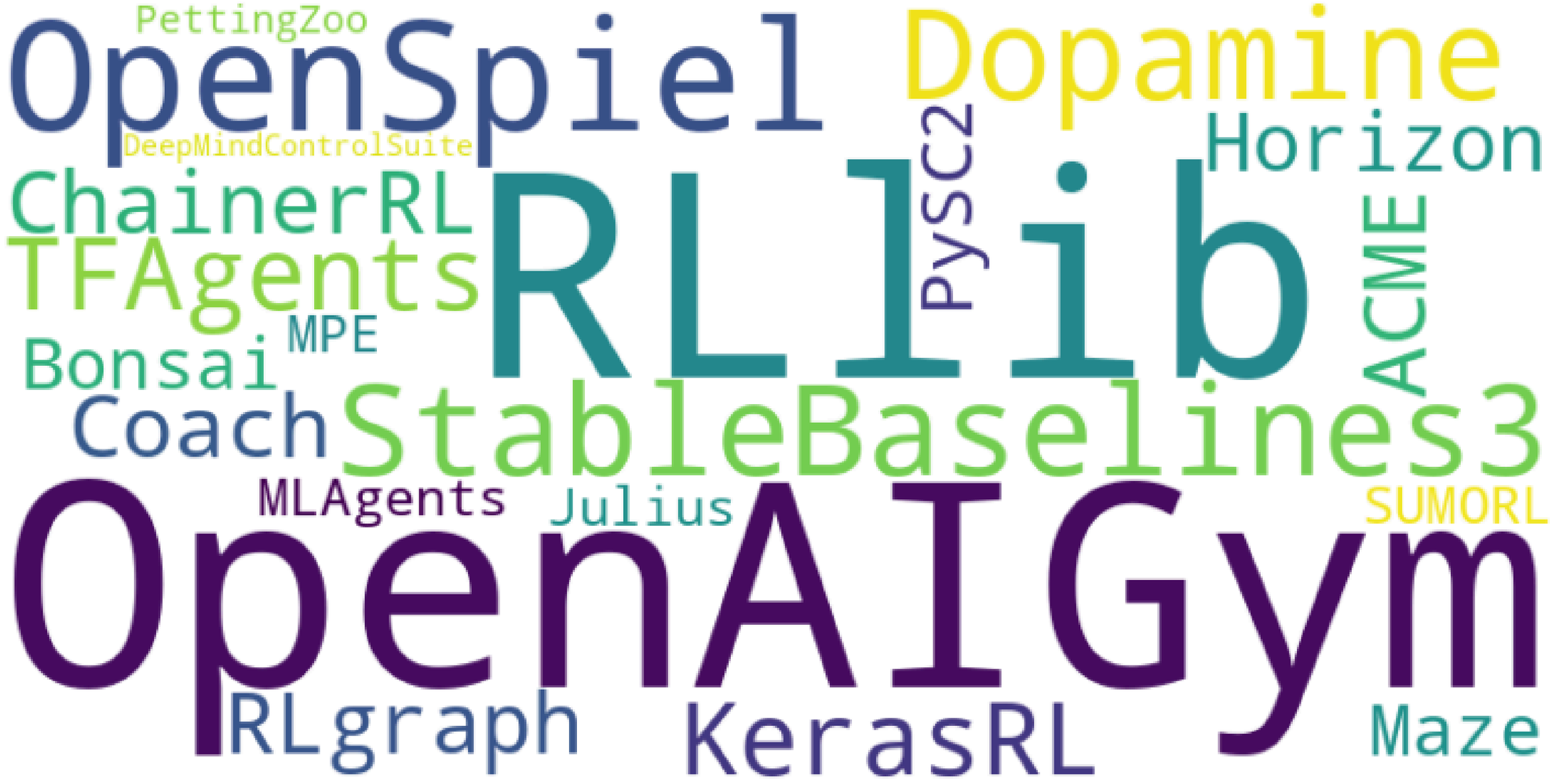
complex policies emerge in RL

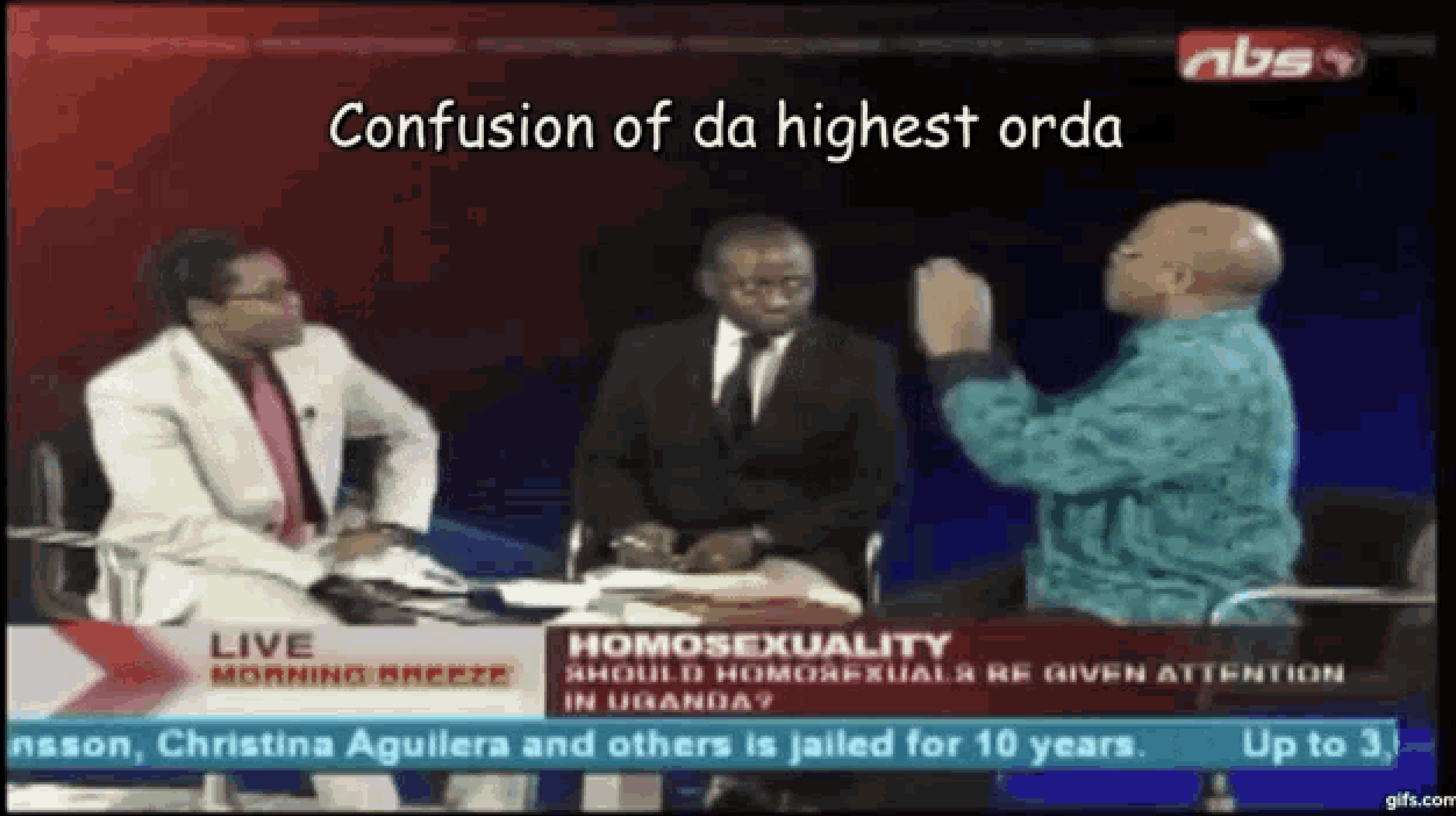
Towards intelligence

RL + DL = AGI

Fine-tuning LLMs - RLHF

- LLMS have to be: helpful, honest, harmless
- Initial Training: Language model trained on vast text data.
- Human Feedback: Humans review model outputs, provide quality ratings.
- Reward Model: Trains to predict human feedback ratings.
- Policy Optimization: Adjusts model to maximize predicted reward.


PettingZoo
OpenSpiel
Dopamine
DeepMindControlSuite
Horizon
ChainerRL
RLlib
PySC2
ACME
TFAgents
MPE
Bonsai
Coach
StableBaselines3
MLAgents
Julius
SUMORL
OpenAIGym
RLgraph
KerasRL
Maze

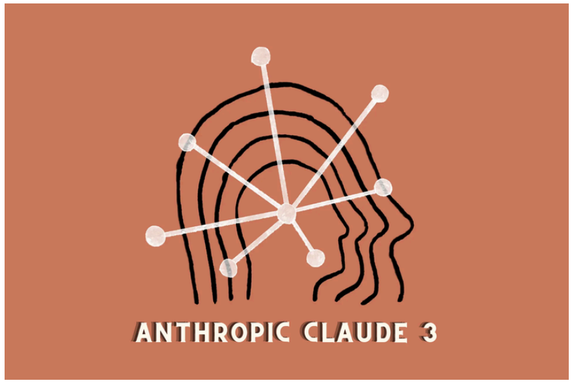

Confusion of da highest orda
nbs
LIVE
MORNING BREEZE
HOMOSEXUALITY
IN UGANDA?
nsson, Christina Aguilera and others is jailed for 10 years.
Up to 3,
gifs.com

Gym-like interfaces

```
import gymnasium as gym

env = gym.make("CartPole-v1")
observation, info = env.reset(seed=42)

for _ in range(1000):
    action = env.action_space.sample()  # todo here goes your action
    observation, reward, terminated, truncated, info = env.step(action)

    if terminated or truncated:
        observation, info = env.reset()
env.close()
```

RL Today

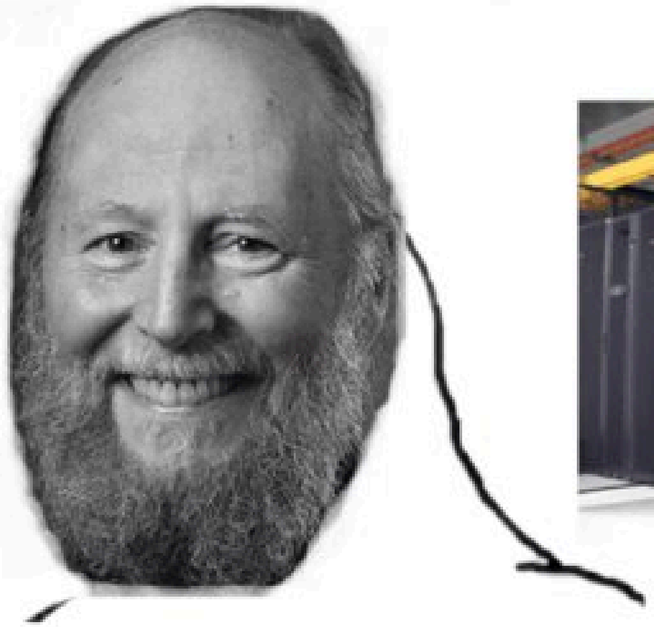
complex
less evident business potential
difficult to interpret
computational needs



Gemini

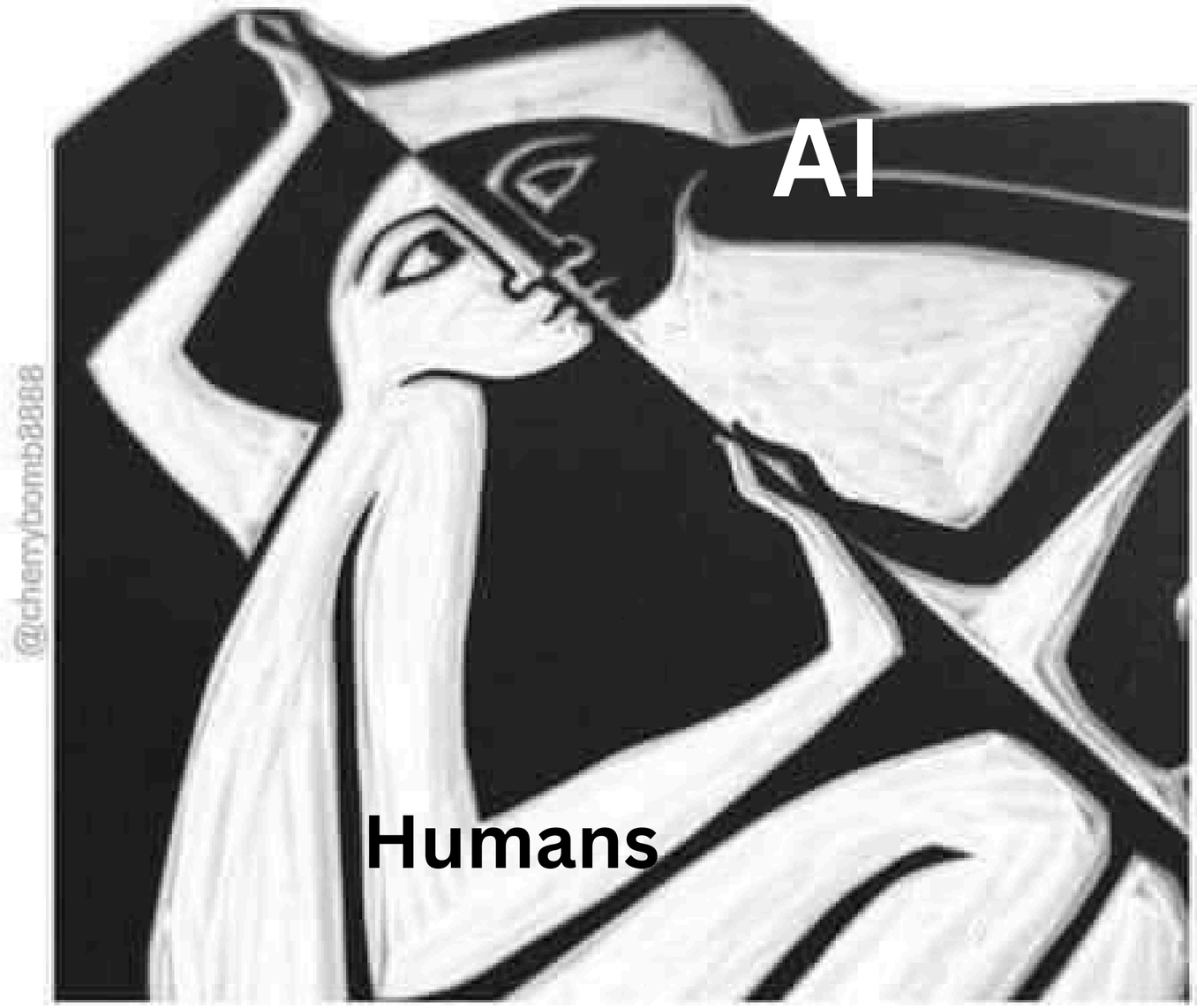
The bitter lesson

nooooo you can't just scale up pure connectionist models on Internet data without inductive biases and modularization and expect them to learn real-world knowledge and grammar from form, or arithmetic and logical reasoning and causal inference—that's just memorization and superficial pattern-matching like Eliza,you need grounding in real-world communication with intent and social dynamics and multimodal robotic embodiment which can foster disentangled learning from guided exploration and self-directed goals expressed in Bayesian programs and probabilistic graphical models which are interpretable and pin down a unique semantics which can be debiased and expressed with uncertainty, and learned efficiently on tiny academic budgets, the cost only shows how this is a dead-end, we need to stop chasing SOTAs and model the complexity of the brain and consider the social context to decolonize AI's structural biases for Third World researchers...



haha gpus go bitterrr

YOUR PERCEPTION OF ME IS A REFLECTION OF YOU;



MY REACTION TO YOU IS AN AWARENESS OF ME

Thank you!

References

David Silver on RL + DL = AGI

Materials from David Silver

Alpha GO

The bitter lesson

OpenAI Gym

Attention is all you need

Classic Control

RLHF