DEEP RL FROM HUMAN PREFERENCES

Paul F Christiano
OpenAI
paul@openai.com

Jan Leike
DeepMind
leike@google.com

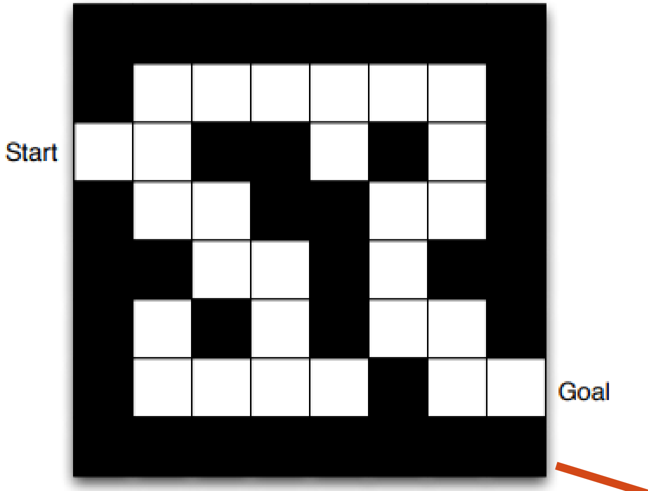
Tom B Brown
nottombrown@gmail.com

Miljan Martic
DeepMind
miljanm@google.com

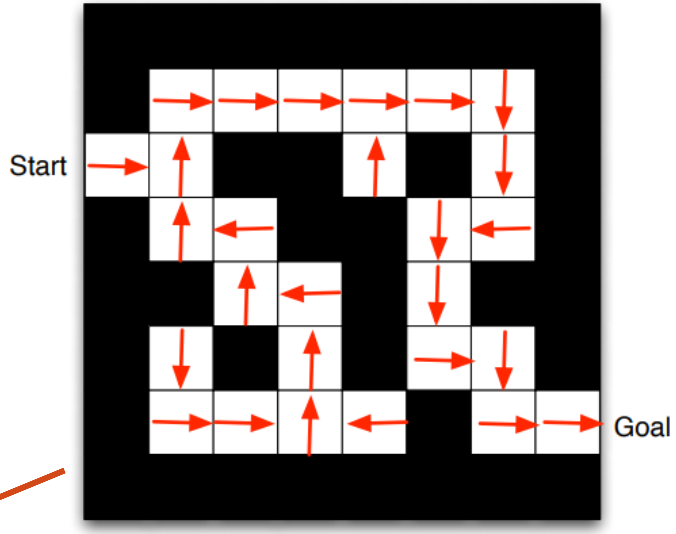
Shane Legg
DeepMind
legg@google.com

Dario Amodei
OpenAI
damodei@openai.com

OVERVIEW



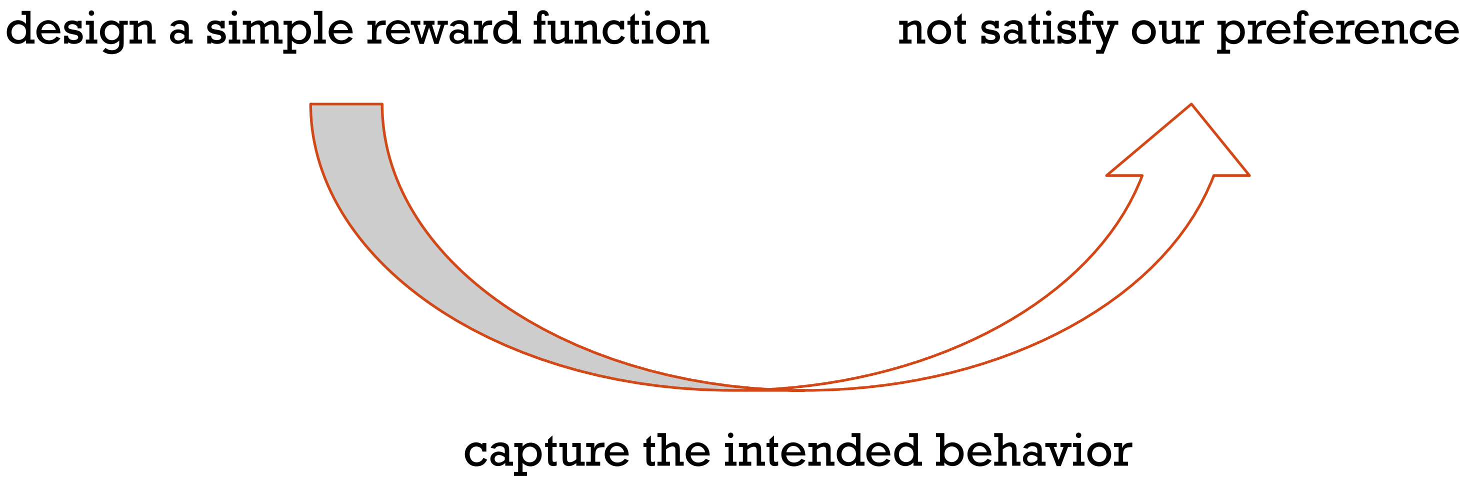
- Rewards: -1 per time-step
- Actions: N, E, S, W
- States: Agent's location



- Arrows represent policy $\pi(s)$ for each state s

RECENT WORK

Many tasks involve goals that are complex and hard to specify



design a simple reward function

not satisfy our preference

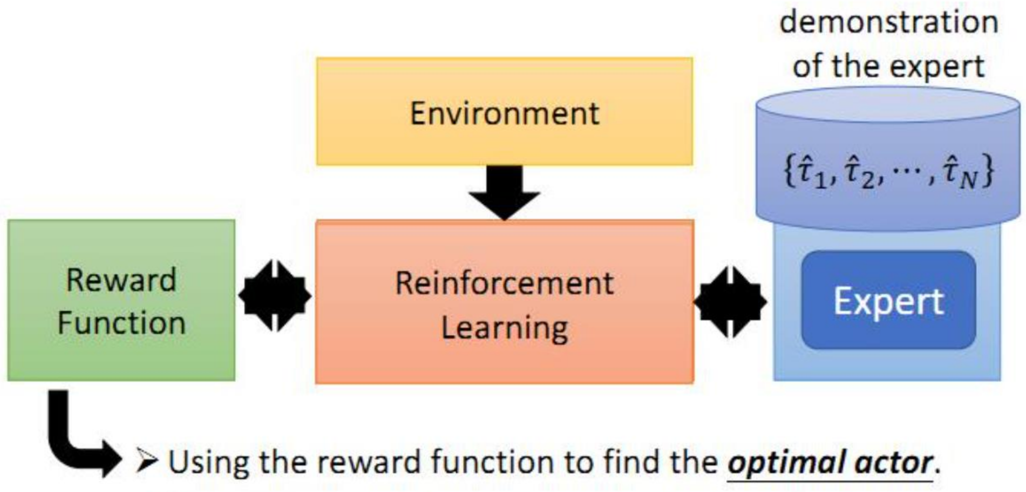
capture the intended behavior

→ **misalignment between the value and human preferences**

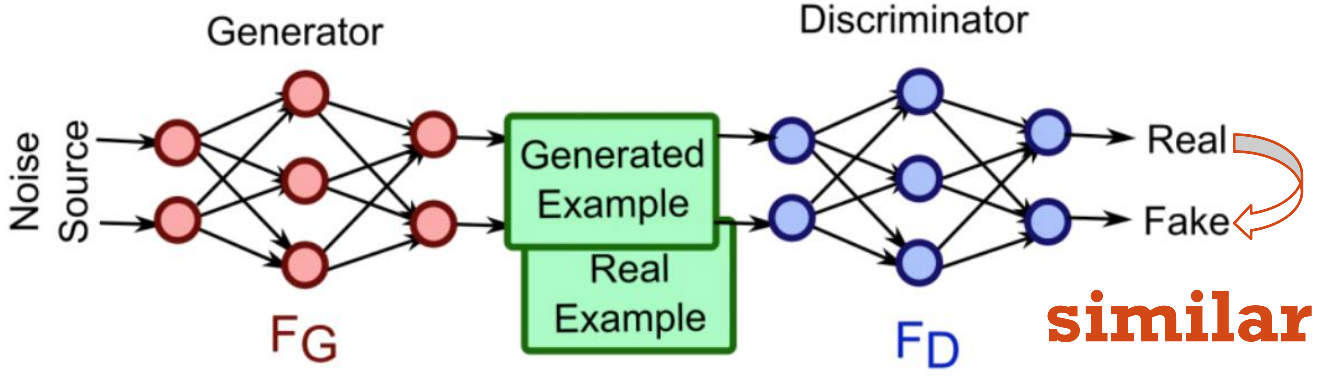
RECENT WORK

The premise : have demonstrations of the desired task

Inverse Reinforcement Learning (IRL)



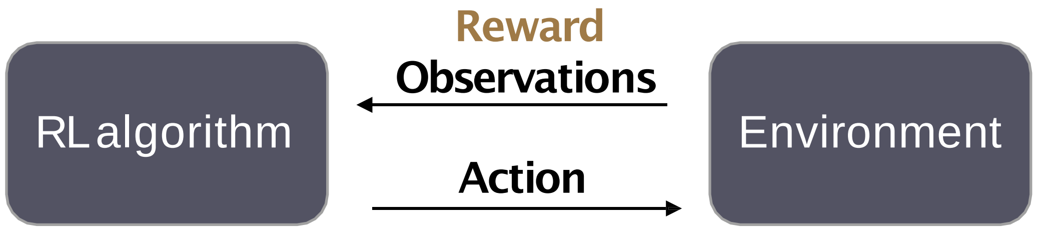
➢ Using the reward function to find the ***optimal actor***.



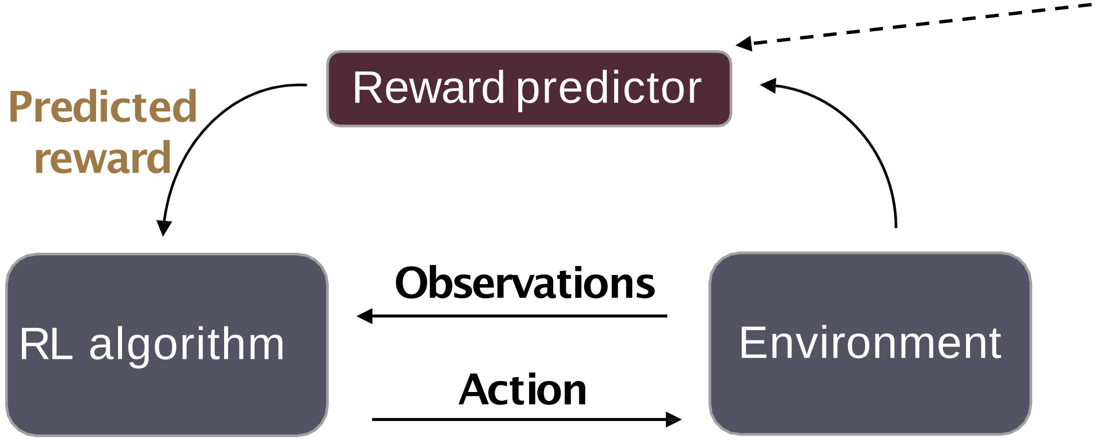
similar

MOTIVATION

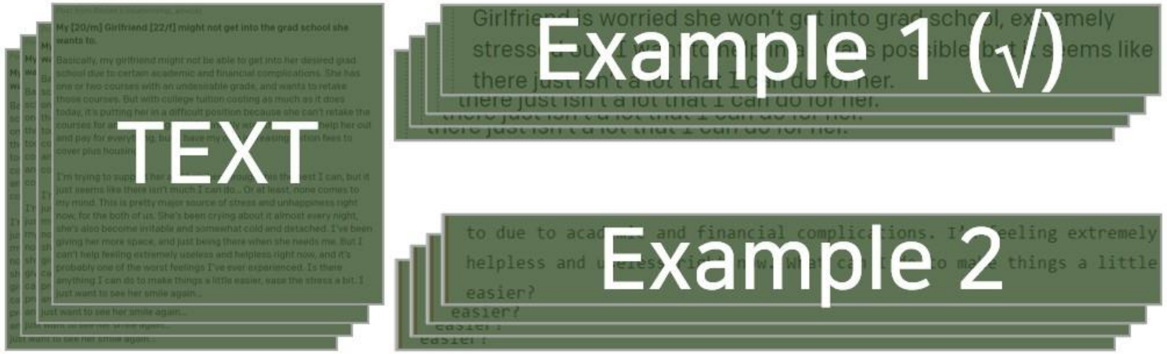
Original RL



RL by human feedback



Set of Comparisons (Human feedback)



Learn a reward function from human feedback and then to optimize that reward function.

Objective

Goal: RL agent produces trajectories which are preferred by the human

while making as few queries as possible to human

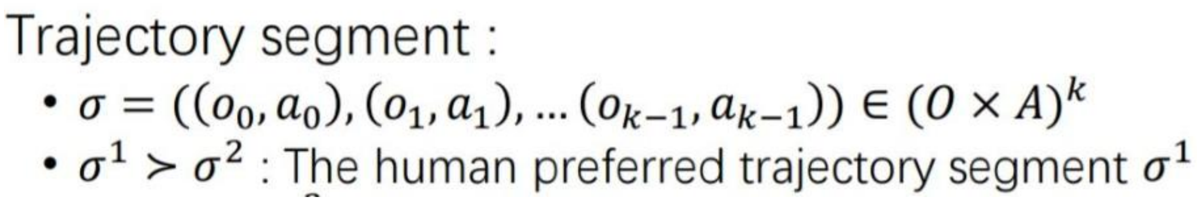
Two neural networks

1. **policy** π **:** $O \rightarrow A$

2. **reward predictor** $r\Sigma$ **:** $O \times A \rightarrow \mathbb{R}$

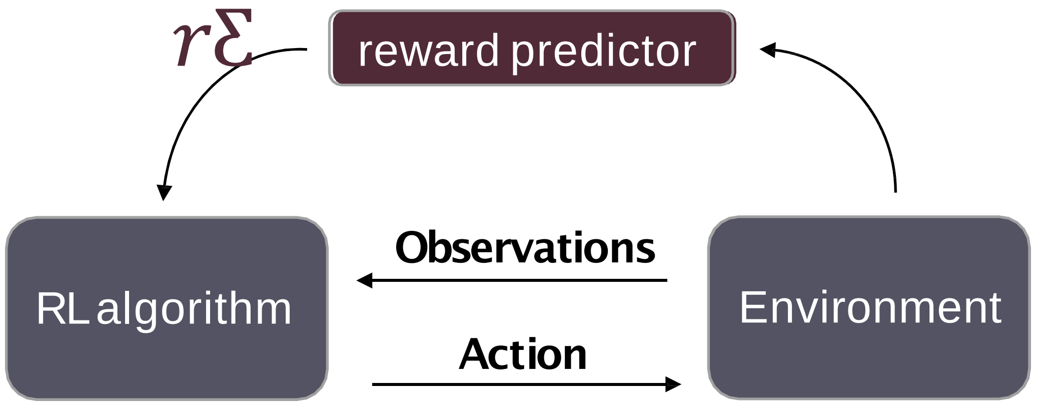
RL agent(policy π**) interacts with the environment to produce trajectories** $\{\tau^1, \ldots, \tau^i\}$**.**

No environment reward



Trajectory segment :

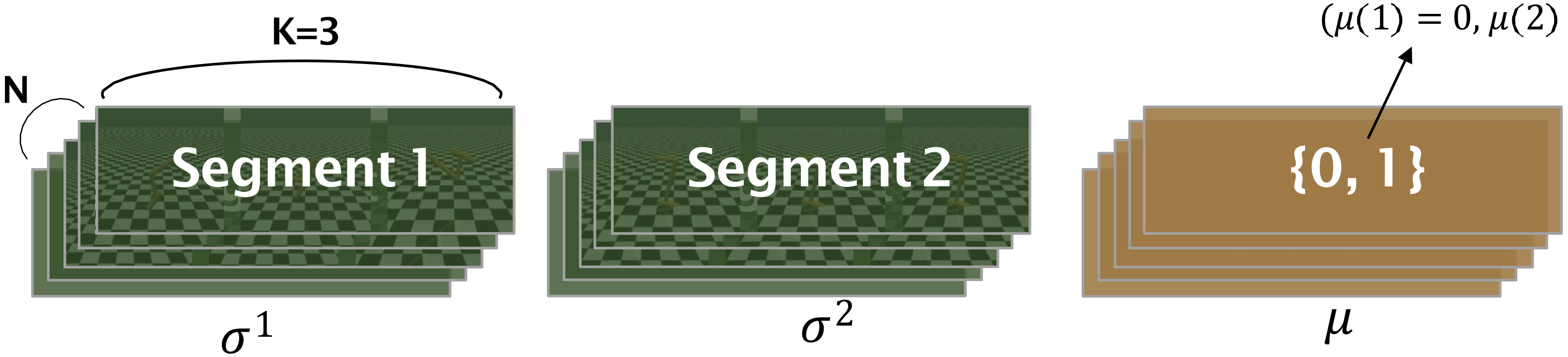
- $\sigma = ((o_0, a_0), (o_1, a_1), \ldots (o_{k-1}, a_{k-1})) \in (O \times A)^k$
- $\sigma^1 \succ \sigma^2$: The human preferred trajectory segment σ^1



Method

Reward predictor r

Human judgements $(\sigma^1, \sigma^2, \mu) \in D$



Segment 2 is preferred
$(\mu(1) = 0, \mu(2) = 1)$

Human's probability of preferring a segment 1

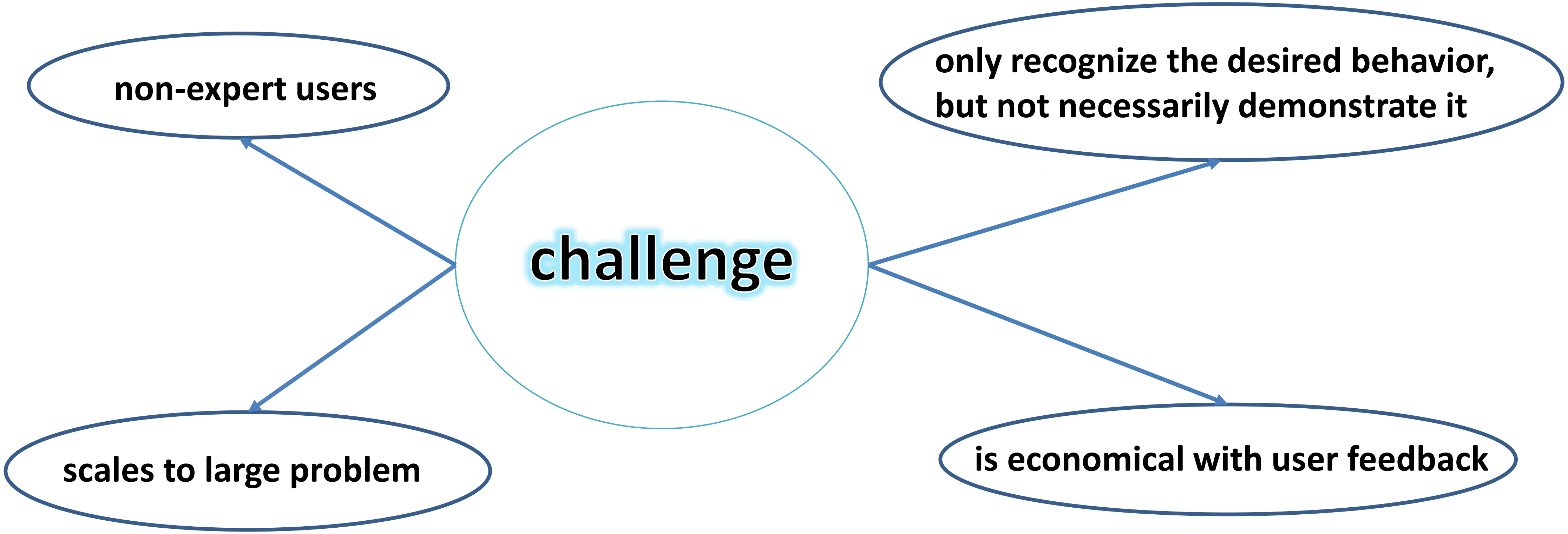
$$\hat{P}(\sigma^1 \succ \sigma^2) = \frac{\exp \sum_{t=1}^{K} \hat{r}(o_t^1, a_t^1)}{\exp \sum_{t=1}^{K} \hat{r}(o_t^1, a_t^1) + \exp \sum_{t=1}^{K} \hat{r}(o_t^2, a_t^2)}$$

Bradley-Terry model

$$loss(\hat{r}) = -\sum_{(\sigma^1, \sigma^2, \mu) \in D} \mu(1) log \hat{P}(\sigma^1 \succ \sigma^2) + \mu(2) log \hat{P}(\sigma^1 \prec \sigma^2)$$

minimize cross-entropy between these predictions($\hat{P}$) and the actual human labels (μ)

Challenge



challenge

non-expert users

scales to large problem

only recognize the desired behavior, but not necessarily demonstrate it

is economical with user feedback

THANKS FOR LISTENING