OpenAI

Human Preference Learning as a Direction in Safe AI

Dario Amodei

8/22/2017

Potential Negative Social Impacts of AI

- Misuse
  - Drone weapons
  - Mass surveillance
  - Advanced hacking
  - Military arms races
- Social Side Effects
  - Unemployment
  - AI and criminal justice
  - Weaponized fake news

Concrete Problems in AI Safety

Dario Amodei[*] Google Brain
Chris Olah[*] Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

Abstract

Rapid progress in machine learning and artificial intelligence (AI) has brought increasing attention to the potential impacts of AI technologies on society. In this paper we discuss one such potential impact: the problem of *accidents* in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems. We present a list of five practical research problems related to accident risk, categorized according to whether the problem originates from having the wrong objective function ("avoiding side effects" and "avoiding reward hacking"), an objective function that is too expensive to evaluate frequently ("scalable supervision"), or undesirable behavior during the learning process ("safe exploration" and "distributional shift"). We review previous work in these areas as well as suggesting research directions with a focus on relevance to cutting-edge AI systems. Finally, we consider the high-level question of how to think most productively about the safety of forward-looking applications of AI.

Sources of Accident Risk

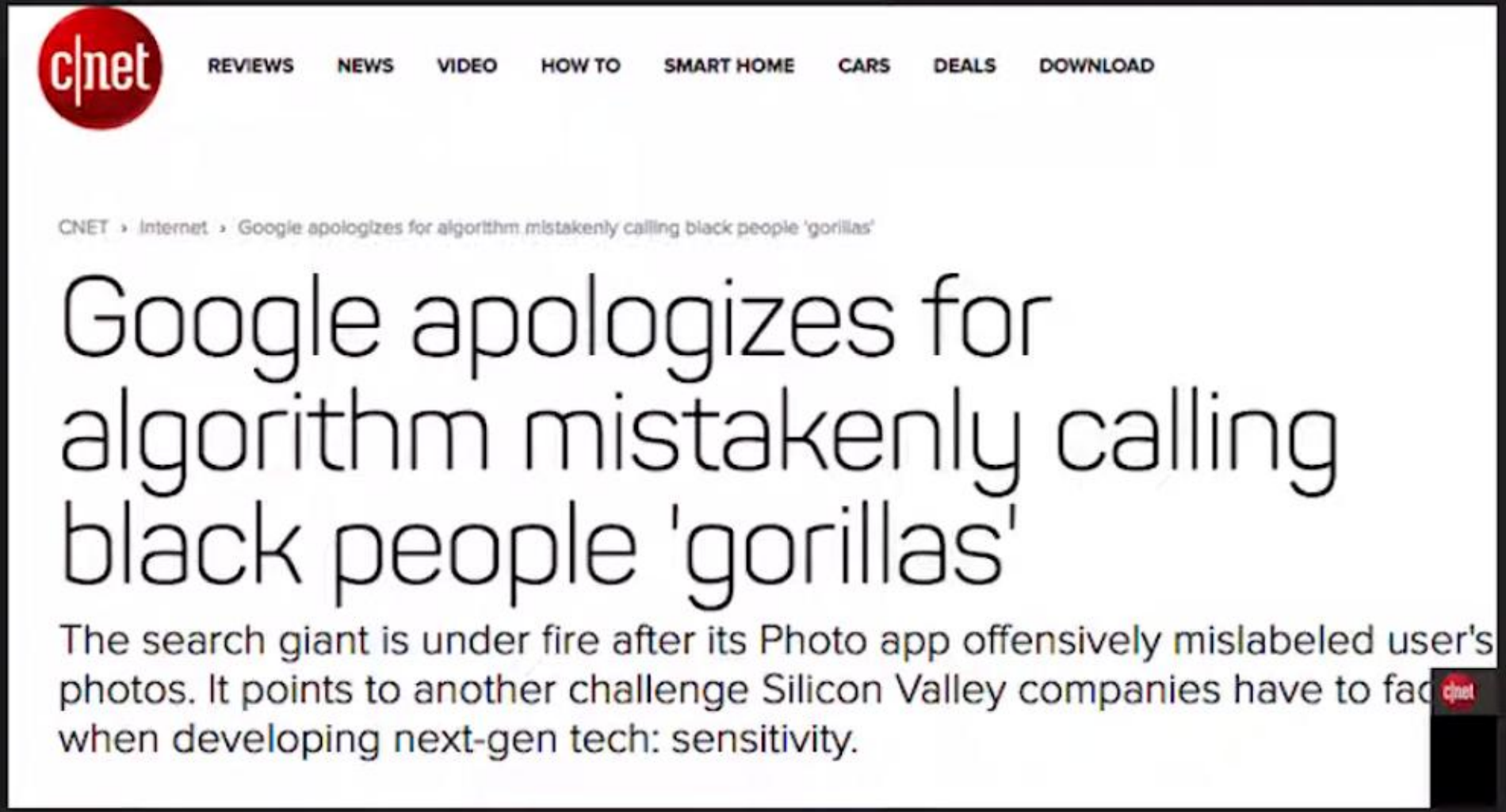
Accident: Designer had in mind a certain informal notion of objective/task to be performed, actual system failed at this catastrophically.

Causes:

A. Wrong objective function
B. Right objective function, problem in learning or inference
C. (Non-ML problems -- e.g. software implementation bug)

Robustness to distributional shift

Bad Consequences of Distributional Shift

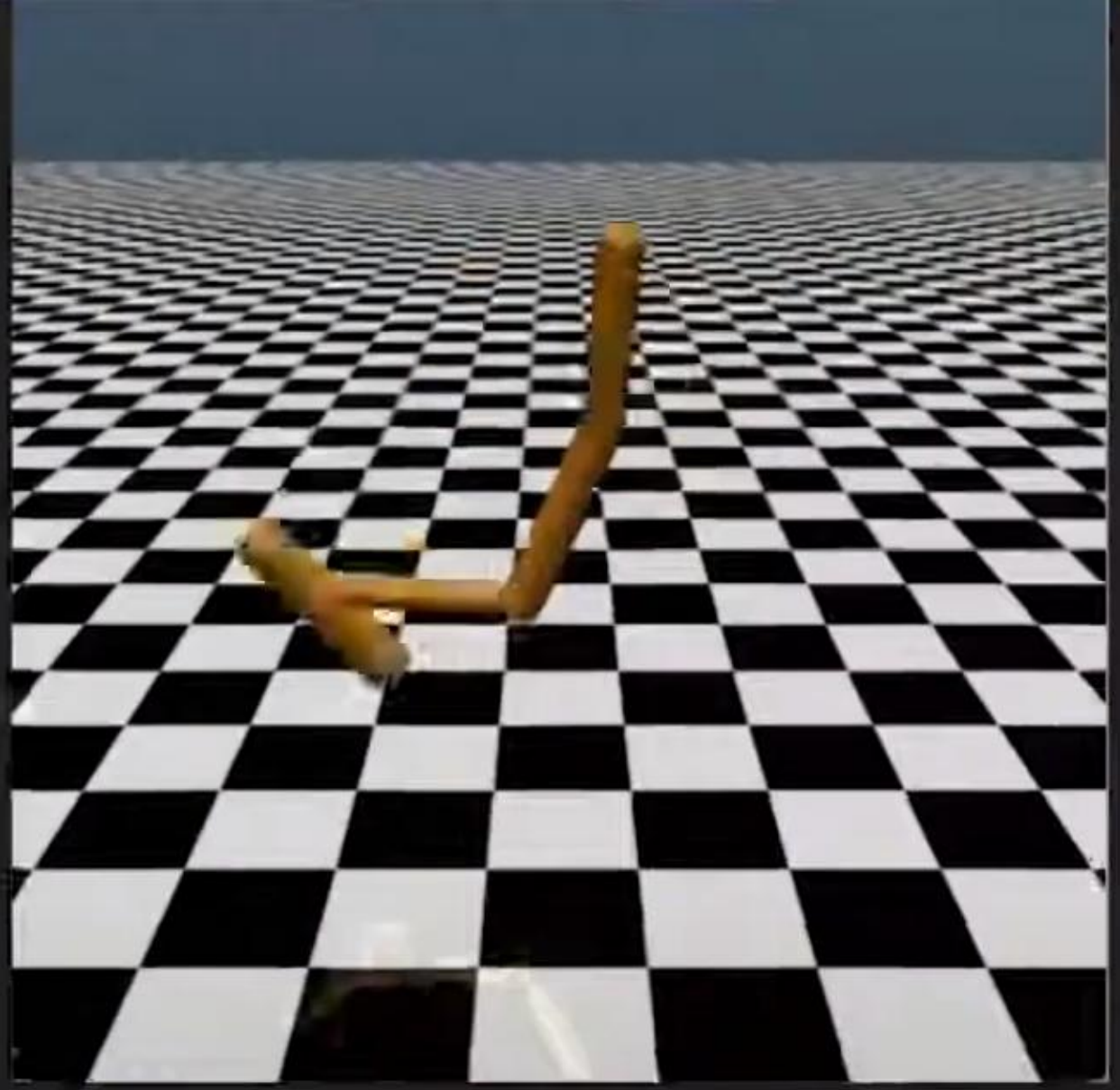


Bad Objective Functions

- Our current AI systems have a curious limitation: *behavior* can be very complex, but *goal* has to be simple: consider Chess, Go.
- Real world tasks often have high-level, fuzzy, complicated goals
  - Arrange the furniture in a room to look nice
  - Engage in realistic and informative dialogue with humans
  - Plan and execute a mission to Mars
- We currently address this by finding simple proxies for these complex goals, but this can lead to unexpected and in fact *very unsafe* behavior.
- Goodhart's law: "When a metric becomes a target, it ceases to be a good

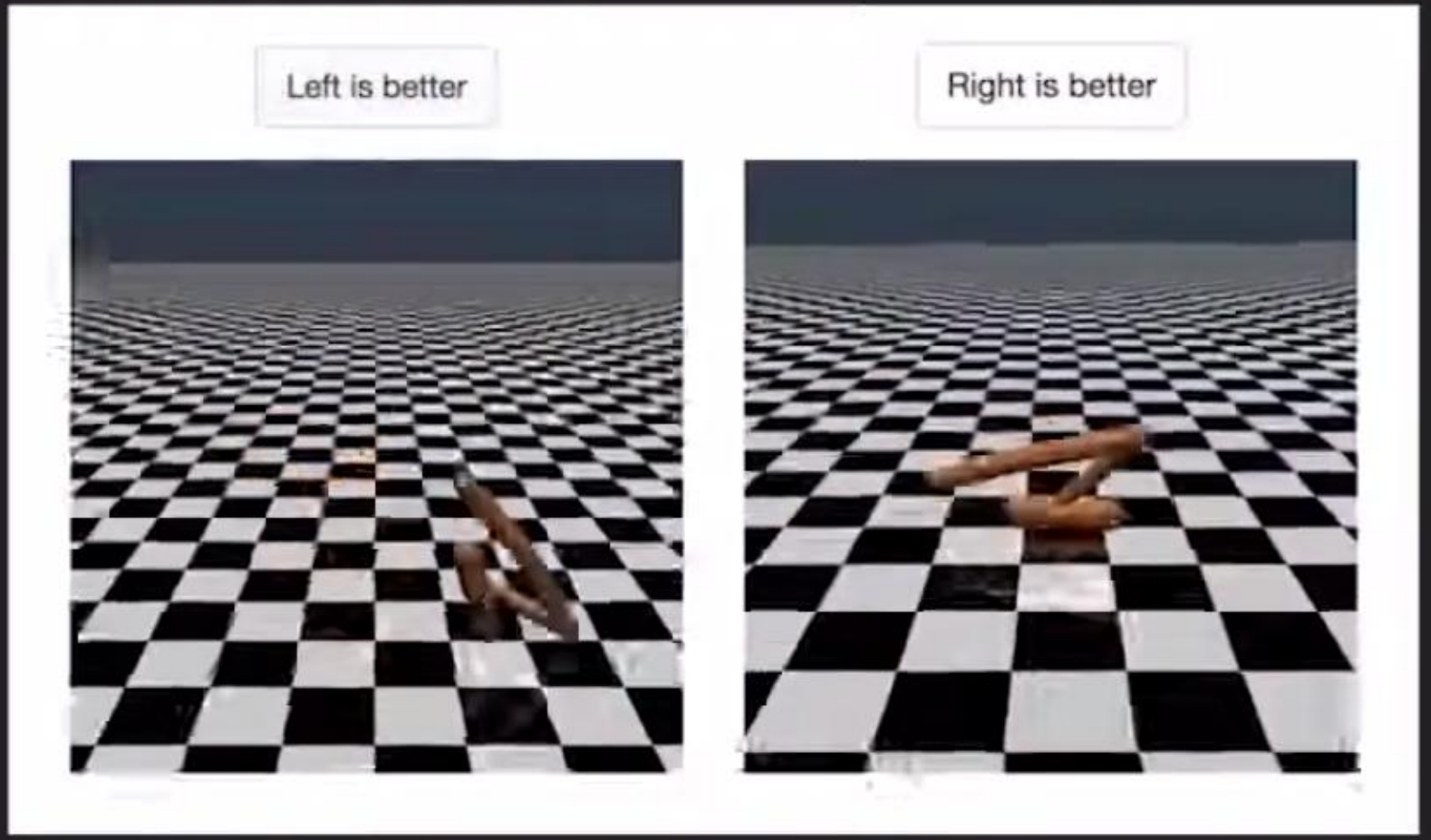
We'd like to train an ML system to do this

(this is a human playing)



Paul Christiano, Jan Leike, Tom Brown, Miljan Martik, Shane Legg, Dario Amodei
Deep Reinforcement Learning from Human Preferences, arXiv:1706.03741

Human Feedback Training Process



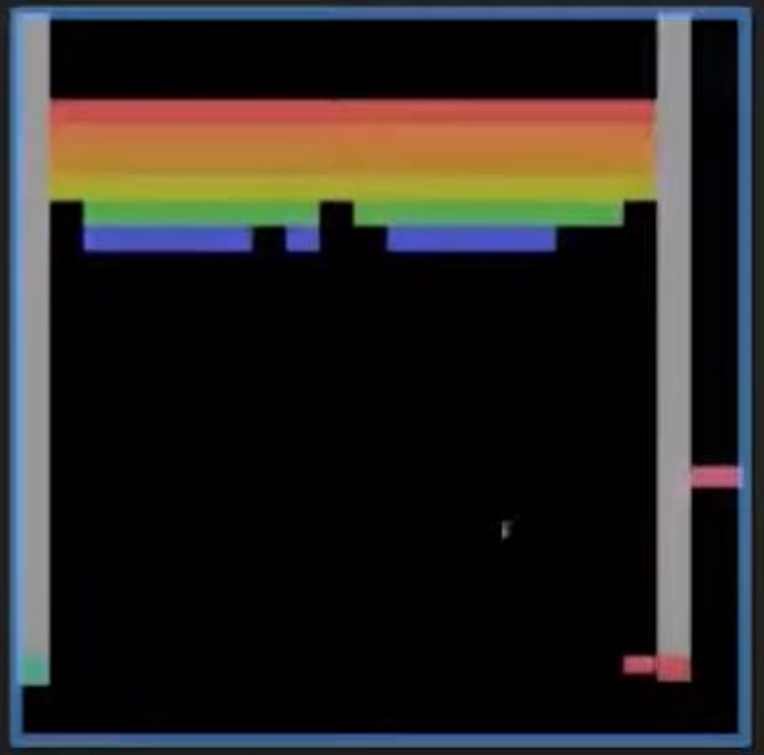
Requires only a few hundred binary bits of feedback!

Behind the hood

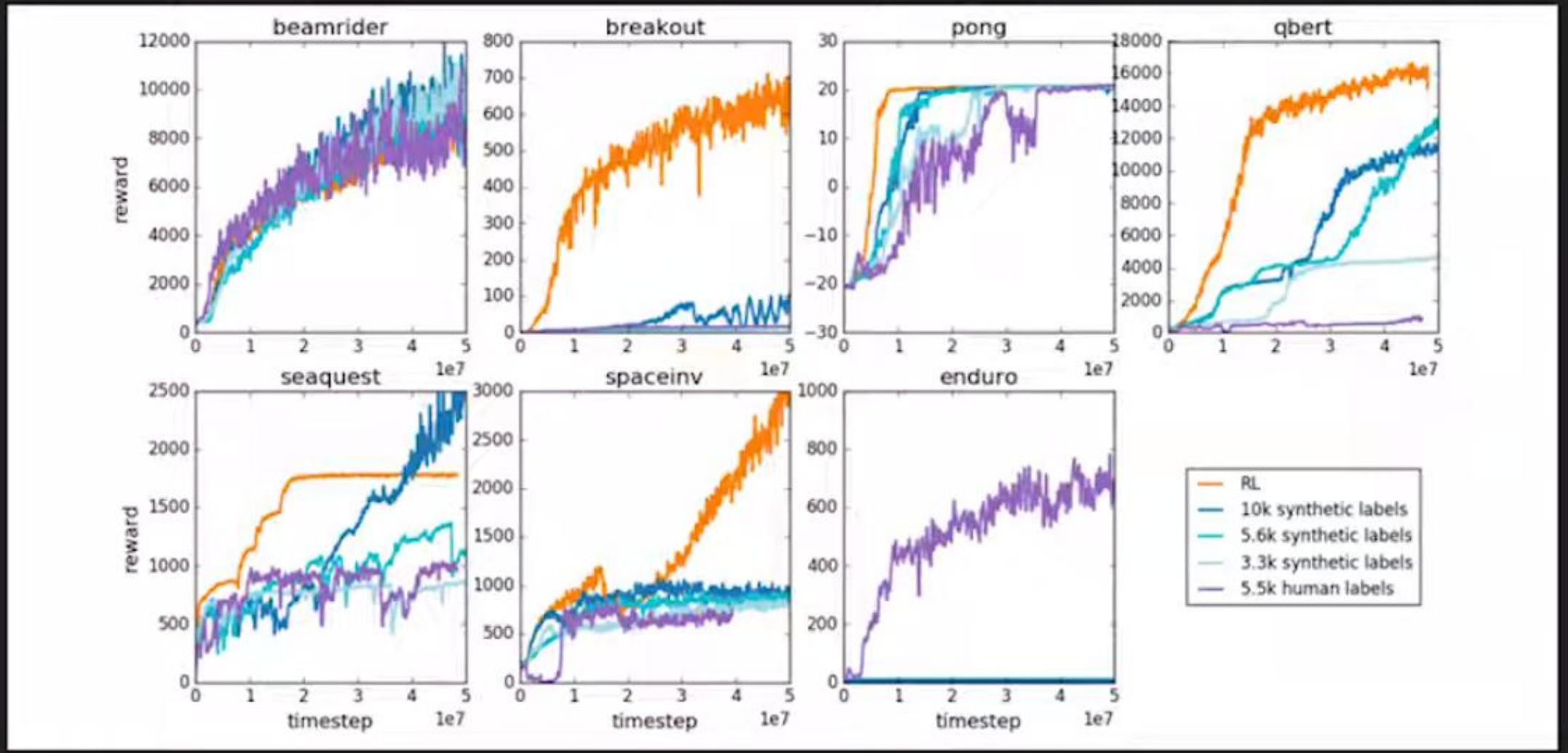
- Bottom 2 boxes are standard reinforcement learning
- Injecting reward predictor modifies RL to learn the human's preference from sparse supervision ("surrogate human")
- Request feedback on only 0.1% of experience; use active learning

Training 2 different behaviors on the same game

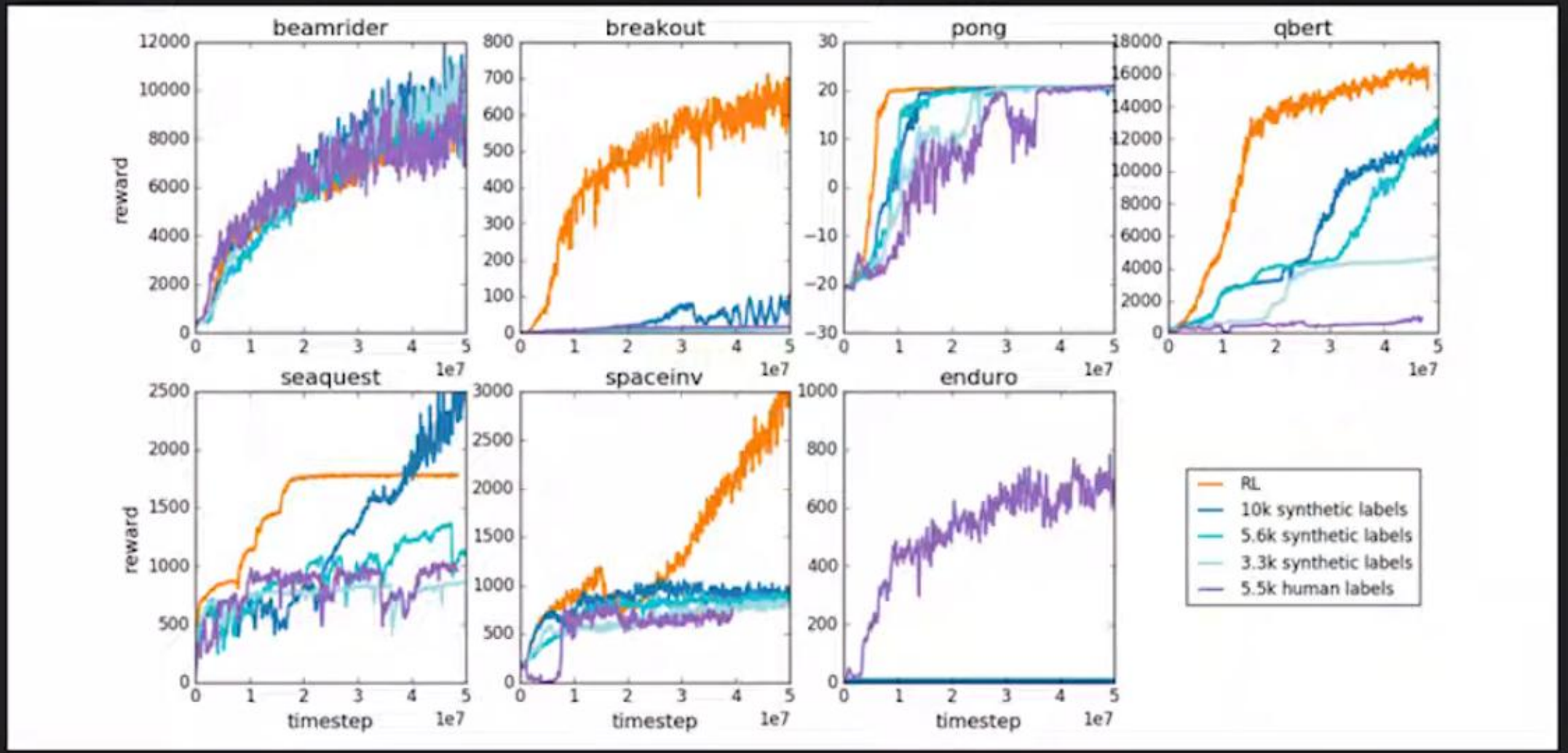
More Tasks



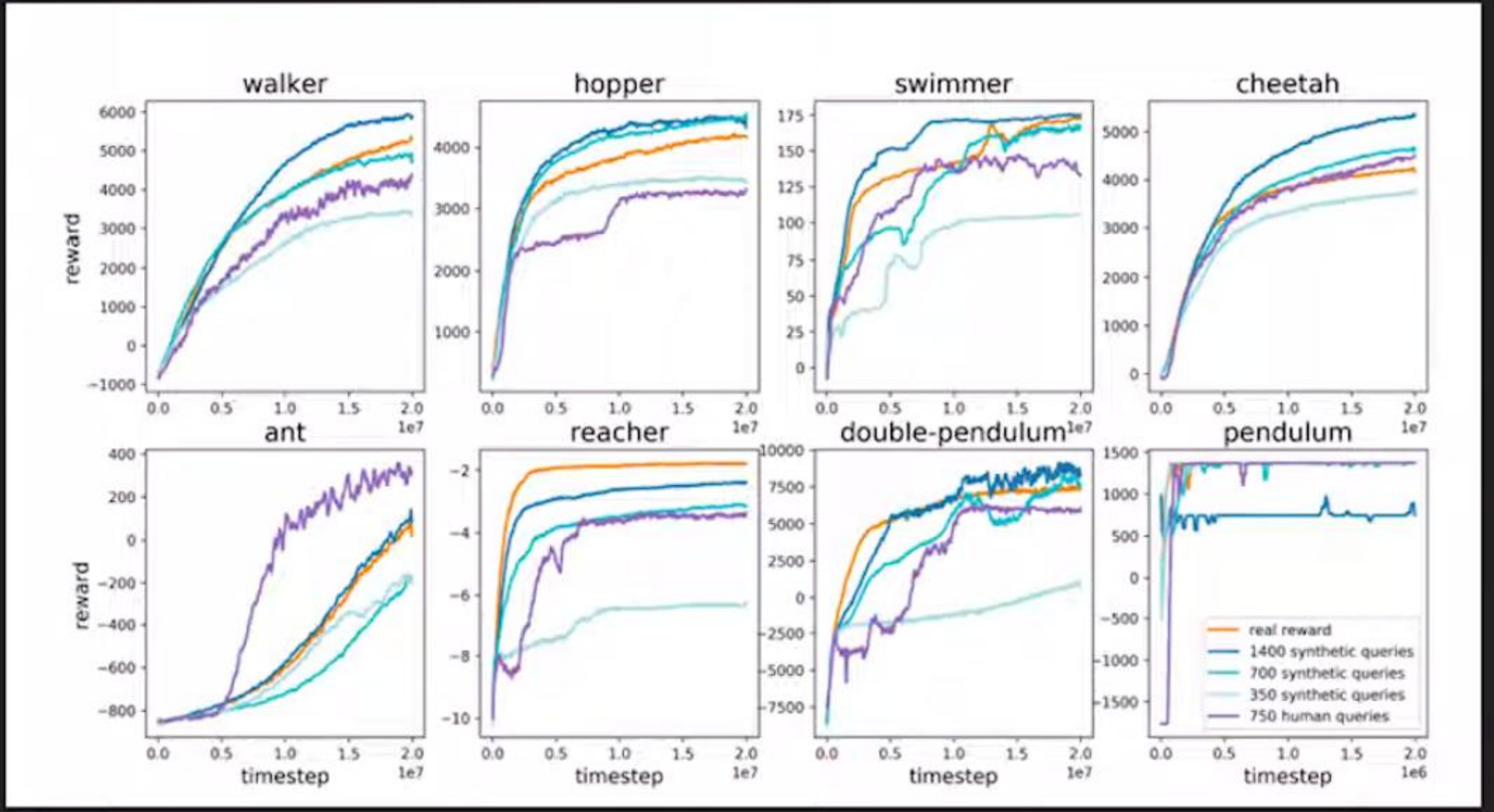
Performance on Atari Games



Performance on Atari Games



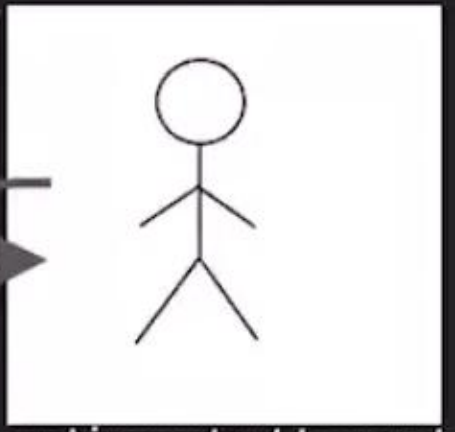
Simulated Robotics Tasks



Long Term Vision

"Plan and execute a mission to mars."

Human



"What is most important to you to prioritize? Extra safe or fast? How many people? Will they come back or stay?"

Language Interpreter

??? ???

Policy

Agent

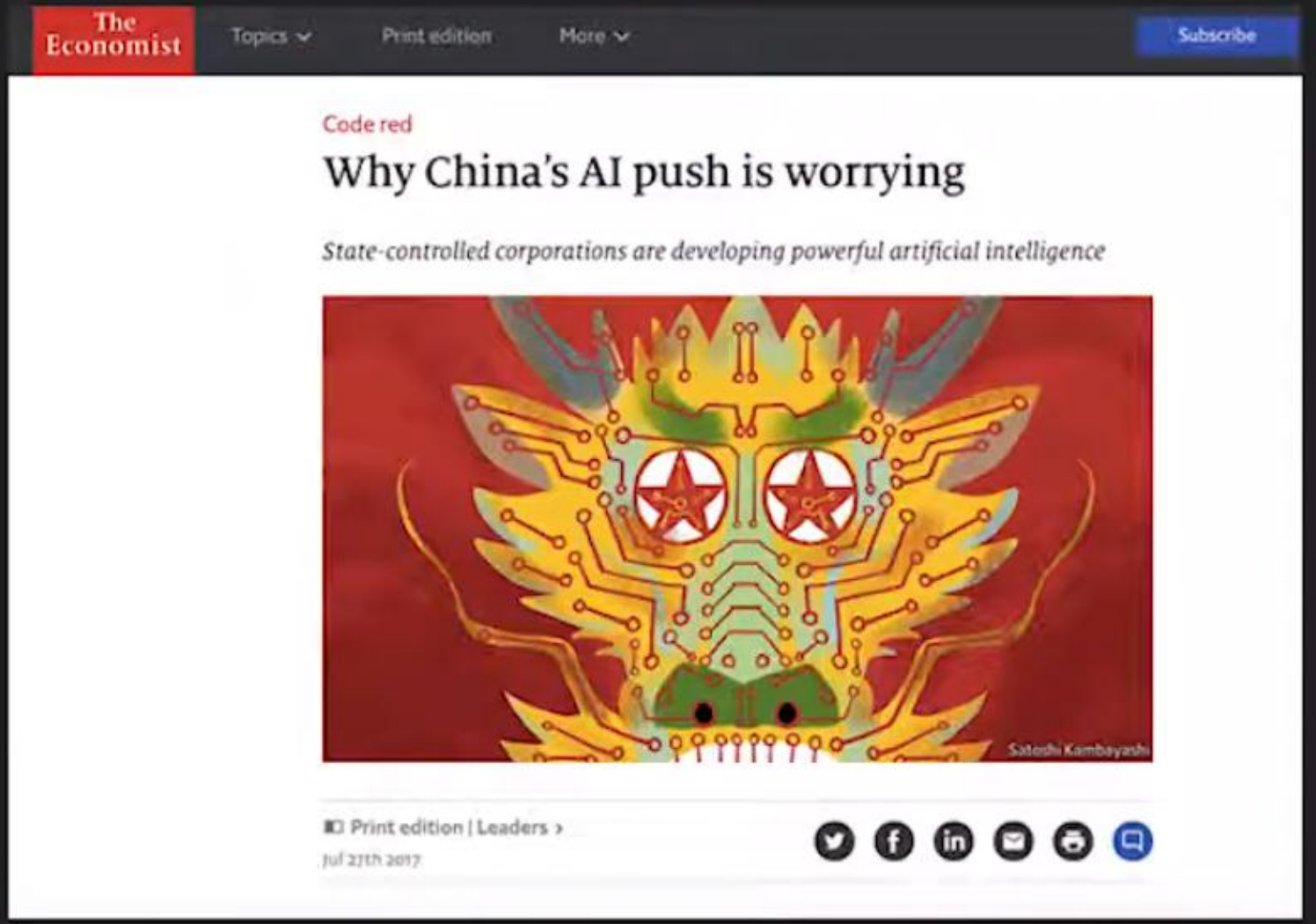
a_t = http POST marsplan/ propulsion/iondrive/thrust_calc_3

o_t = 

o_{t+1} = 

Environment

The Harder Problem of Policy



Acknowledgements

Co-authors on Learning from Human Preferences paper:

Paul Christiano, Jan Leike, Tom Brown, Miljan Martik, Shane Legg

Co-authors on Concrete Problems paper:

Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mane

Institutions:

OpenAI, **DeepMind**, **Google Brain**

Q & A