Center for Human-Compatible Artificial Intelligence

Center for Human-Compatible AI
University of California, Berkeley
Berkeley Way West
Berkeley, CA 94709
humancompatible.ai

Mapping the Political Economy of Reinforcement Learning Systems—Spring Semester

Sponsored by the Center for Human-Compatible AI at UC Berkeley, and with support from the Simons Institute and the Center for Long-Term Cybersecurity, we are convening a cross-disciplinary group of researchers to examine the near-term policy concerns of Reinforcement Learning (RL). RL is a rapidly growing branch of AI research, with the capacity to learn to exploit our dynamic behavior in real time. From YouTube's recommendation algorithm to post-surgery opioid prescriptions, RL algorithms are poised to permeate our daily lives. The ability of the RL system to tease out behavioral responses, and the human experimentation inherent to its learning, motivate a range of crucial policy questions about RL's societal implications that are distinct from those addressed in the literature on other branches of Machine Learning (ML).

We began addressing these issues as part of last semester's Simons Institute program on the Theory of Reinforcement Learning. This semester we would like to broaden the discussion to include perspectives from Law and Policy. The aim of this working group will be to establish a common language around the state of the art of RL across key societal domains. From this examination, we hope to identify specific interpretive gaps that can be elaborated or filled by members of our community. Our ultimate goal will be to map near-term societal concerns and indicate possible cross-disciplinary avenues towards addressing them.

Among the questions we will be exploring:

- How do existing regulations influence the adoption of RL across particular domains, such as transportation, social media, healthcare, energy infrastructure? What distinctive forms of regulation are missing?

- How do formal assumptions and domain-specific features translate into particular algorithmic learning procedures? For example, how does reinforcement learning "interpret" domains differently than supervised and unsupervised learning, and where does this matter from a policy standpoint?
- How do the assumptions and design decisions (e.g. multi-agency, actively shaping dynamics vs. modeling them) behind RL systems differ from those of cyber-physical systems? What are the implications of these differences with respect to regulation and proper oversight of these systems?

All materials will be organized by Thomas Krendl Gilbert, Ph.D. candidate in Machine Ethics and Epistemology at UC Berkeley. Those interested are invited to email Thomas (tg340@berkeley.edu) if they wish to be added to the list of participants and provide input on readings and or discussion topics.

Zoom links for meetings are available on request.

Spring 2021 Reading Plan + Schedule

The tentative schedule going forward will be **Mondays 12pm-1pm pacific time**. Meetings will be **weekly**, alternating between discussion and new reading materials. Also, meetings **will be recorded** and audio will be shared with participants afterward. Topics, papers, and discussants listed below will be finalized throughout the semester as participants see fit.

2/15: Introductions, brief presentation on last semester, brainstorm for particular papers / topic areas / invited speakers

2/22: What is the Political Economy of AI?

- Can we automate tax policy using RL?
- Code is Law: On Liberty in Cyberspace (Lawrence Lessig)

3/1: Political Economy of AI (cont'd)

- Political Economy of Machine Translation (Steven Weber)

3/8: Towards a Political Economy of RL

- Mapping the Political Economy of RL Systems (Thomas Krendl Gilbert)

3/15: Measurement and Evaluation in Content Recommendation

- From optimizing engagement to measuring value (Milli et al)

3/22: Deep RL—What is it and why does it matter?

- Societal Implications of Deep Reinforcement Learning (Whittlestone et al)

3/29: Content Recommendation, Continued

- Smitha Milli Follow-up Discussion

4/5: RL in Healthcare: App Recommendations

- Yonatan Mintz (relevant papers are here, here, and here)

4/12: Differential Games

- David Fridovich-Keil (Probably the most important one is here, and two application papers I'll talk about are here and here. I'll also mention this one briefly.

4/19: What do people think are the best ways of broadening PERLS and publishing work?

- Open discussion (no papers planned)

4/26: RL in Traffic: Benchmarking and Managing Behaviors

- Eugene Vinitsky

5/3: Energy Infrastructure

- Roel Dobbe

5/10: Bandits & Markets

- Lydia Liu (two-sided markets)

5/17: Workshop planning for NeurIPS 2021!

5/24: School Assignment

- Sam Robertson