Literature overview

Richard Möhn

19th February 2020

Contents

1 To search

- What has Paul written about RL-based IDA?
- How are rewards determined in RL?
- Since I mention IL, read something about IL?
- Find something about how ML algorithms deal with faulty data/outliers. – Noisy data with little repetition. – Mathematical function learning with noise. – Detecting/classification with mislabelled data (credits: Logan Smith).
- DONE See todos in SupAmp-ReAmp and Overfail2
- DONE Search backward from Christiano et al. (2018) on Google Scholar
- Something about surrogate modelling? It appears to be related with distillation.

2 Potential sources

- Possibly relevant works citing Christiano et al. (2018), according to Google Scholar:
  - Backward search on Google Scholar

- Modeling AGI Safety Frameworks with Causal Influence Diagrams
- Evolutionary Computation and AI Safety: Research Problems Impeding Routine and Safe Real-world Application of Evolution
- Multiparty Dynamics and Failure Modes for Machine Learning and Artificial Intelligence
- Risks from Learned Optimization in Advanced Machine Learning Systems

3 To skim and decide

- Resources from Christiano et al. (2018) that I've marked with a blue cross.
- Semi-supervised reinforcement learning
- Scalable agent alignment via reward modeling: a research direction

4 To (re-)read

- https://www.lesswrong.com/posts/fq7Ehb2oWwXtZic8S/reinforcement-learning-in-the-iterated-amplification
- Christiano 2016b
- Christiano 2019

5 Annotated bibliography

> Often the assembling of an annotated bibliography is a distinct stage in a research process [...]. Each annotation is an opportunity to evaluate the credibility of a source, summarize its argument, and explain its relevance to your project.
>
> [...] If you can't summarize your sources or explain their relevance, you are likely not ready to write your paper. (Booth et al. 2016, p. 102 f.)

Paul Christiano (2016a). *Reliability amplification*. URL: https://www.alignmentforum.org/posts/6fMvGoyy3kgnonRNM/reliability-amplification (visited on 2019-09-02) TODO: Copy summary from notes and clean up. Add relevance.

Paul Christiano, Buck Shlegeris and Dario Amodei (2018). 'Supervising strong learners by amplifying weak experts'. In: arXiv: 1810.08575 [cs.LG] TODO: Copy summary from notes and clean up. Add relevance.

References

Booth, Wayne C. et al. (2016). *The Craft of Research*. 4th ed. The University of Chicago Press. DOI: 10.7208/chicago/9780226239873.001.0001.

Christiano, Paul (2016b). *The reward engineering problem*. URL: https://www.alignmentforum.org/s/EmDuGeRw749sD3GKd/p/4nZRzoGTqg8xy5rr8.

– (2019). *Thoughts on reward engineering*. URL: https://www.alignmentforum.org/posts/NtX7LKhCXMW2vjWx6/thoughts-on-reward-engineering (visited on 2019-08-30).