Draft Basis

What I need for planning the Farlamp draft

Richard Möhn

19th February 2020

(For a project overview and a glossary, see the home page of the Farlamp repository.) The questions in the following are adapted from Booth et al. (2016, p. 175). Some of the section headings quote the chapter titles in that book.

1 Readers

Who are my readers?

- The members of LessWrong and the AI Alignment Forum. Perhaps the members of MIRIxDiscord.
- Ideally conference or workshop attendants, or readers of a journal. I don't know if I can get my article into such circles, though.

What do they know?

- Most know ML and CS better than I.
- They might not know about IDA.

Why should they care about my problem? IDA is a major approach to AI alignment. How reliable it is, we don't know, and therefore not whether and what precautions are needed. My research would provide empirical evidence to help answer these questions.

2 Ethos

What kind of ethos or character do I want to project? From Booth et al. (2016, p. 119):

- '[support] claims with evidence that readers accept'

- '[consider] issues from all sides'
- 'anticipate and address [readers'] questions and concerns'
- 'thoughtfully [consider] other points of view'
- 'acknowledge other views and explain [my] principles of reasoning in warrants'

→ 'give readers good reason to work *with* [me] in developing and testing new ideas'

3 Question and answer

Sketch my question and its answer in two or three sentences. Question: Can SupAmp or ReAmp stay reliable despite overseer failure? The answer is to be determined.

4 Reasons and evidence

Sketch the reasons and evidence supporting my claim. – TBD

5 Acknowledgements and responses

- What questions, alternatives and objections are my readers likely to raise?
- How do I respond to them?
- Christiano 2016c talks about policies being capability-amplified. That sounds like they are the assistants. So shouldn't the project be about assistant failure, rather than overseer failure?

  Response: It depends on what 'amplify' means, which appears to vary. – In Christiano 2016a it sounds like the overseer is getting amplified, in Christiano 2016b it sounds like the assistant is getting amplified, and in Cotra 2018 both the overseer and the assistant are arguments to the `Amplify` procedure.

  In the end it doesn't matter for this project. The overseer is a simple procedure that will fail if it gets wrong input from an assistant. So I might as well stick with what I've written so far and situate failures in the overseer.
- Christiano 2016c assumes assistants that are powerful enough to be able to negotiate with each other. In IDA generally the first assistant is trained to almost human level. (This might be different with low bandwidth overseers (Saunders 2018); I haven't thought it through.) So for distillation we need powerful learning algorithms, which don't exist yet. And in Overseer failures in SupAmp and ReAmp I predict that failure tolerance depends on the learning algorithm. Then won't the results of this project be irrelevant in a few years (months?), when we have different learning algorithms?

Response: I'm not aiming to document the response of specific learning algorithms to overseer failure. Rather, I'm testing the hypothesis that failure tolerance depends on various parameters, such as learning algorithm and regularization strength. And I'm testing the hypothesis that only when the overseer failure rate exceeds a certain threshold (the failure tolerance of a particular configuration) does the overall failure rate blow up. These hypotheses should hold independent of the learning algorithm used.

To be continued.

6 Warrants

- When may my readers not see the relevance of a reason to a claim?
- Can I state the warrant that connects them?

TBD

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 (2016a). *ALBA: An explicit proposal for aligned AI*. URL: https://ai-alignment.com/alba-an-explicit-proposal-for-aligned-ai-17a55f60bbcf (visited on 2020-01-03).

– (2016b). *Capability amplification*. URL: https://ai-alignment.com/policy-amplification-6a70cbee4f34?#.wmeq2iqwv (visited on 2020-01-03).

– (2016c). *Reliability amplification*. URL: https://www.alignmentforum.org/posts/6fMvGoyy3kgnonRNM/reliability-amplification (visited on 2019-09-02).

Cotra, Ajeya (2018). *Iterated Distillation and Amplification*. URL: https://www.alignmentforum.org/posts/HqLxuZ4LhaFhmAHWk/iterated-distillation-and-amplification-1 (visited on 2019-09-06).

Saunders, William (2018). *Understanding Iterated Distillation and Amplification: Claims and Oversight*. URL: https://www.greaterwrong.com/posts/yxzrKb2vFXRkwndQ4/understanding-iterated-distillation-and-amplification-claims (visited on 2020-01-03).