Draft Basis

What I need for planning the Farlamp draft

Richard Möhn

13th September 2019

(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?

TBD

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.