A Rational Analysis of Persuasion

S. A. Barnett

1 Background

Identify the topic your project will explore, and briefly provide some of the context for your project, describing what previous research has found in this area.

- Topic to explore: Models of debate - specifically, involving two (or more) debaters competing in an adversarial context, and one judge who must make a decision on the strength of the debates.
- Context for project:
  - The RSA model for incorporating theory of mind.
  - Work on self-play in games.
  - The weak evidence effect.
  - 'Debate' as an approach for iterative, aligned improvement in AI systems.
- Previous research: The Gricean pragmatic explanation for the weak evidence effect. Order statistics and sticks?

2 Question

State the specific question you are going to examine in your final project. Does the weak evidence effect improve judgment?

3 Method

Briefly describe the method you are going to use to try to answer this question. Give some of the details behind your experimental procedure, your approach to modeling or analyzing the data, your plans for analyzing the model, or the position you will take in your review.

- Experiment 1: Testing humans for the weak evidence effect in the MNIST game, using OpenAI code.

- Experiment 2: Testing for the weak evidence effect in the judge for the MNIST debate game. Namely: Does weak but supportive evidence in one direction cause a shift in probability mass in the other direction?
- Experiment 3: Allow the judge in the MNIST debate game to accommodate for the weak evidence effect by modifying the judge to be an RSA-style 'pragmatic listener'. Contrast the difference in overall performance accuracy with that of the original judge.

References

[1] Craig R M McKenzie, Susanna M Lee, and Karen K Chen. "When Negative Evidence Increases Con®dence: Change in Belief After Hearing Two Sides of a Dispute". In: *Journal of Behavioral Decision Making* 15.1 (2002), p. 17.

[2] Philip M. Fernbach, Adam Darlow, and Steven A. Sloman. "When good evidence goes bad: The weak evidence effect in judgment and decision-making". In: *Cognition* 119.3 (June 2011), pp. 459–467. ISSN: 00100277. DOI: 10.1016/j.cognition.2011.01.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010027711000394 (visited on 10/15/2019).

[3] Geoffrey Irving, Paul Christiano, and Dario Amodei. "AI safety via debate". In: *arXiv:1805.00899 [cs, stat]* (May 2, 2018). arXiv: 1805.00899. URL: http://arxiv.org/abs/1805.00899 (visited on 05/08/2019).