Aron Vallinder

Preferences: Elicitation, Extrapolation, and Aggregation

Abstract: This dissertation considers ”full information” theories of value, especially in the context of machine superintelligence. The discussion is structured around three main questions: (i) How do we determine an agent’s preferences? (ii) How do we extrapolate current preferences to ones that would be had were full information available? (iii) If such fully informed preferences are not consistent across a group of agents, how do we aggregate the individual preferences? In answering the first question, I survey empirical data and theoretical work in economics, neuroscience and preference elicitation. As for the second question, I attempt to solve some of the standard problems that arise: How do we characterize ”full information”? In what sense can the fully informed preferences be said to be the same as those previously held? Apart from answering these questions, I also look at potential methods for extrapolating preferences. These methods are evaluated from both an empirical perspective (Are they computationally and practically feasible?) and an ethical perspective (Are the results of such methods desirable?) For the third question, I consider work in preference and judgment aggregation, with particular focus on how to measure the distance between different rankings. The extrapolation methods vary in how coherent different resulting preferences are, and I investigate whether there is a unique aggregation procedure that is feasible for all of these methods.

Thesis outline

1. Introduction

    a. ”Full information” theories of value
    b. Relevance for machine superintelligence

2. Determining preferences

    a. Revealed preferences in economics
    b. Insights from neuroeconomics
    c. Preference elicitation in machine learning

3. Extrapolation

    a. What is ”full information”?
    b. Previous extrapolation algorithms
    c. Some new classes of extrapolation algorithms
    d. Evaluating extrapolation algorithms

4. Aggregating extrapolated preferences

    a. Previous work in preference and judgement aggregation
    b. Some novel suggestions
    c. Matching extrapolation and aggregation methods

5. Conclusions

Aron Vallinder is enrolled at Lund University,
concurrently pursuing an MA in Theoretical Philosophy, and a BSc in Engineering Mathematics.