SuperBrain

Roger Bishop Jones

7 January 2016

Contents

1 Proposal **3**

I The Analytic Superbrain **4**

2 Introduction **5**
2.1 Method 5
2.2 Problems 6

II The Holistic Superbrain **7**

Chapter 1

Proposal

The proposed work is a monograph of uncertain length.

The character of the work will be philosophical, broadly within the tradition which has recently been known as *analytic* philosophy, but which arguably extends back to Socrates and beyond.

There will be a single overarching subject matter through which some of the fundamental problems of analytic philosophy will be motivated. That subject is the relationship between people and machines as machines become ever more capable of undertaking intellectual as well as physical tasks, and as global networking facilitates collaboration between peopla and information processing machinery. Among the questions considered are questions of trust and of authority in this context.

In this matter the distinction between analytic and other propositions is particularly significant, and is reflected in the structure of the work.

A provisional title for the work is *SuperBrain*. It will consist of one introductory chapter followed by two parts, *The Analytic SuperBrain* and *The Holistic Superbrain*.

The introductory chapter will have the following structure:

- Background
- The Problem
- The Method
- The Analytic

Part I

The Analytic Superbrain

Chapter 2

Introduction

In this introduction I want to give a first explanation of the project in hand, and the methods to be used. To do that, I think I need to give a little background.

The project in hand concerns *AI*, Artificial Intelligence. It involves architectural design for a globally distributed intelligence, the components of which include both people and engineered artefacts (computers and networks). If you think about the design of intelligent artefacts at a sufficiently high level, in a sufficiently systematic way, it is hard to avoid addressing and taking a position on some philosophical problems. If you are unlucky you may find, as I have, that the most natural explanations of sound engineering principles can only be given in the teeth of opposing conventional wisdom from contemporary professional philosophers. If you are a philosophically minded engineer, these philosophical problems may assume a greater interest than the more worldly aspects of the enterprise, and architectural design may become a thought experiment around which a philosophical enterprise is built. This is what has happened to me. This work has resulted.

In my thinking about intelligence I have been interested primarily in problems which can be precisely formulated, and which have (if any) definite solutions which can be reliably checked. Most mathematical problems have these characteristics, which carry forward to suitably well formulated applications of mathematics. Many problems of engineering design are reducible to crisp problems using mathematical models of the problem domain, and so we can think of the target domain as encompassing design automation.

There is a particular reason for separating out these crisp problems and for considering their automation in a different way to the way we consider the automation of less crisp AI problems, like natural language processing and many other aspects of the problem of building artefacts which emulate, simulate or replicate human behaviour. That reason is trust.

In the literature of science fiction the spectre of an artificial intelligence running amok recurrs. There are long standing principles, introduced by Asimov, guiding the design of Robots, which are intended to ensure their proper behaviour. But other parts of the literature recognise that artificial intelligence, just like todays viruses and worms, may not observe any benign laws. They may be designed with malicious intent, they may simply emerged or evolved from processes unknown to us and beyond out control, or they be the product of benign incompetence. *The Singularity* to chose one name for such inhuman intelligence, is generally expected to be quite beyond our control.

2.1 Method

This work is intended primarily as *analytic philosophy*. In this section I will describe some aspects of the kind of philosophical analysis to be employed.

- Attitude towards natural/ordinary lan-

- Methods from mathematics.

2.2 Problems

Part II

The Holistic Superbrain