Why Foundations Matter

Roger Bishop Jones

Contents

Mathematical Foundations and Existential Risk

1 Introduction

The foundations of mathematics, perhaps even more so than mathematics itself, are a closed book to most people who are not themselves mathematicians or philosophers. They might nevertheless readily accept that the foundations of mathematics are *important* on the basis of knowing that mathematics is essential to science and engineering, the products of which underpin our wealth and well-being.

Though accepting that *if* mathematics has "foundations" then they *must* be important, many would have no idea what such a foundation might be. They might be surprised to discover that there are many alternative foundations, and considerable controversy about which, if any, should be adopted. Such controversy may be philosophical or pragmatic, and may involve philosophers, mathematicians, and/or computer scientists.

If then we pressed the non-specialist for an opionion on whether the choice of mathematical foundation system mattered, he might be in doubt.

There is reason to believe that the foundations of mathematics, and possibly even the choice of system from among the alternatives, might be much more important than the above informal argument might lead us to suppose. It is my purpose here to facilitate a discussion aimed at making this seem intelligible to the non-specialist.

2 Discussion Plan

The plan is for a philosophical discussion which illuminates the issues, so the principal elements in the plan are topics for discussion. and the principal purpose of the notes is to provide background for the discussions.

I propose that the discussion be in three main parts:

1. an introduction to logical foundations for mathematics

   The main idea here is to get an informal sense of the kind of thing that is called a "foundation for mathematics". This is mainly through a brief exposition, partly historical, with any discussion which might arise from that.

2. superintelligence and existential risk

   A brief introduction to the work of Nick Bostrom, and his recent book on the above topic.

3. logical foundations for safety

   The possible relevance of mathematical foundation systems to the implementation and application of superintelligence while minimising existentialmrisk

Of these the first and last may be primarily expository, the greatest time being devoted to a general informal discussion of the central topic.

3 Logical Foundations for Mathematics

3.1 What is a Foundation for Mathematics

When mathematicians and philosophers talk of a "foundation system for mathematics" they are usually talking about a formal logical system, in which mathematicians can define mathematical concepts and prove theorems about these concepts. This idea is due to Gottlob Frege, is also associated with Bertrand Russell whose "Principia Mathematica" was the first substantial consistent formal derivation of mathematics using such a logical system.

Informally this can be understood using the example of axiomatic set theory. In order to be able to define mathematical concepts it is necessary to have available an ontology of abstract objects which can be used to model the required mathematical structures. An ontology of pure sets suffices for this purpose. (pure sets are sets built up from the empty set, so that only sets are needed). The sets serve rather like lego bricks, you combine them together to create more elaborate sets of arbitrary complexity which serve as numbers, or matrices or mathematical fields (for example). Once mathematics is built up in this way, scientists and engineers can use the mathematics to construct mathematical models of various parts of aspects of the physical world so that we can predict behaviour and design buildings and machines which fulfill various purposes.

3.2 Conference on Foundations

There will be a symposium at Birkbeck College London in january on "different approaches to the foundations of mathematics".

> The focus of this conference is on different approaches to the foundations of mathematics. The interaction between set-theoretic and category-theoretic foundations has had significant philosophical impact, and represents a shift in attitudes towards the philosophy of mathematics. This conference will bring together leading scholars in these areas to showcase contemporary philosophical research on different approaches to the foundations of mathematics. To accomplish this, the conference has the following general aims and objectives. First, to bring to a wider philosophical audience the different approaches that one can take to the foundations of mathematics. Second, to elucidate the pressing issues of meaning and truth that turn on these different approaches. And third, to address philosophical questions concerning the need for a foundation of mathematics, and whether or not either of these approaches can provide the necessary foundation.

A central theme will be the comparative merits of two "foundation systems", the well established set theoretic foundations, and a more recent system called "homotopy type theory" (HoTT). The former is based on the concept of set, a concept simple enough that for a while it was taught to primary school pupils in the United Kingdom. The latter is based on the concept of "weak omega-groupoid", which comes from algebraic topologic and higher order category theory, two of the most abstract branches of mathematics which are mostly studied only by postgraduate students or postdoctoral researchers.

HoTT is very fashionable at present because the core idea comes from a "Fields medalist" (the most famous and coveted prize for mathematicians, similar to a nobel prize) who departed from his core competence in Algbraic Topology to make proposals about the foundations of mathematics (the province of mathematical logicians) it has been developed by an interdisciplinary group of academics including mathematicians, computer scientists and philosophers.

Are these discussions merely academic, or do these issues have any importance to ordinary people? In the next section I describe an issue which might have enormous significance, if not for us, for our children, and then argue that these two problems, the foundations for mathemathematics, and "existential risk" are connected.

4 Superintelligence and Existential Risk

Oxford now has an interdisciplinary "Future of Humanity Institute" directed by Nick Bostrom, also a Professor of Philosophy at the University of Oxford. His latest book on "superintelligence" is creating a stir.

4.1 Elon Musk Tweets about AI risks

Elon Musk is a prominent technocrat and entreprener, best known as the founder of Tessla the luxury electric car manufacturer and SpaceX (space exploration technologies corp). Musks ideas are taken seriously by many, and he has recently commented (though Twitter) on the risks in AI.

> Worth reading Superintelligence by Bostrom. We need to be super careful with AI. Potentially more dangerous than nukes.
>
> Hope we're not just the biological boot loader for digital superintelligence. Unfortunately, that is increasingly probable.

The book he refers to is written by an Oxford Philosophy professor, and is concerned primarily with the "existential risk" arising from a future "singularity" in the development of artificial intelligence, a point at which the pace of development becomes very fast indeed and results in superintelligences intellectually superior to homo-sapiens which take over the dominant role in our planetary ecosystem, subjugating and possibly exterminating humanity.

Elon Musk is not the first to take these risks seriously, many writers over the decades since the development of digital computers have anticipated machine intelligence and been concerned with its risks.

4.2 The Singularity

Growth is linear (straight line) when the same increase occurs in each unit of time. It is exponential when the same percentage increase occurs in each unit of time (e.g. 5(colloquially it is now used to mean very fast or explosive, but strictly speaking that is not what it means). An example of this is represented by advances in miniturisation and power of semiconductors, described by "Moore's Law" which states that every 18 months the density of transistors in leading edge semiconductor fabrications will double. Many other examples are found in biological populations, epidemiology.

A singularity is a point at which a curve "goes to infinity", which will happen if the same amount of growth occurs in an ever decreasing unit of time.

Observers have noticed that the speed of technological change is ever increasing, and have calculated that there will if past trends are continued, be a singularity in the graph of technological capability some time early in the 21st century. Usually in this contect the idea of a singularity refers to the point at which a "superintelligence" is created, it being assumed that such a being will be capable of rapidly designing and even more intelligent being and the intelligence will explode "exponentially".

4.3 Existential Risk

Bostrom's book "SuperIntelligence, ..." is devoted primarily to "existential risk". This concerns the possibility common in science fiction but now often discussed as a medium term real risk, that a superintelligent machine will go rogue and will first take over control from human beings, and ultimately allow our species to go extinct.

How credible do we think this risk is, and what can we do about it?

5 Mitigating Existential Risk

There are many ways in which the various risks discussed by Bostrom might be eliminated or mitigated. We are here concerned only with the connection with "logical foundations".

I have done some work on an extended essay entitled "Positive Philosophy and The Automation of Reason" which has a bearing upon this. This connects with an approach to the automation of reason which has its roots in Leibniz, and involves the use of calculating machines (today, digital computers) orovided with a logically encoded store of knowledge of the laws of mathematics and science, and is able to compute reliably the answer to any questions put to it. Leibniz's idea is clearly a forerunner of todays ideas about artificial intelligence, but delivers intelligence (or superintelligence) moderated by the following two factors:

1. The machine has no power to do other than answer questions put to it.
2. The machine has no "motivation", it cannot be regarded a friendly or antagonistic, to humanity, it just answers questions with complete accuracy (provided its knowledge has been correctly set up).

Leibniz's ideas preceded the invention of logical foundations for mathematics, but if implemented today could be be implemented using such a logical system. This would guarantee certain features of the system which further mitagates risks (though in this case, risk of error rather than of malfeasance). This is one of the reasons why I have been advocating that approaches to this kind of artificial intelligence should be based on logical foundation systems.

5.1 Weaknesses in Bostrom's Discussion?

On a first partial reading it seems to me that Bostrom's thinking in this area suffers from two pervasive difficulties.

The first is that, despite his seeing that intelligence is orthogonal to motivation (that we should not assume that an intelligent artefact will share our values or objectives) he still assumes that they will all be autonomously motivated.

The second is that he seems to neglect consideration of the social control mechanisms which we have for controlling ordinary intelligent (possibly malevolent) beings. Partly this may be because he presumes that a "super" intelligence will find it easy to circumvent this kind of control, just as merely human hackers are able to do. However, this is to assume an assymmetry between good and evil. Superintelligence will be applied to enhancing the security of our networks, and the various other control systems which prevent anyone from taking over control of the whole universe.

On the other hand, the existence now and in the future of malevolent humans, ensures that once artificial intelligence is here, it will be applied for nefarious purposes. A future conflict between good and evil, largely played out in cyberspace, does seem highly probable, as an escalation of the present day standoff between malware and security software.