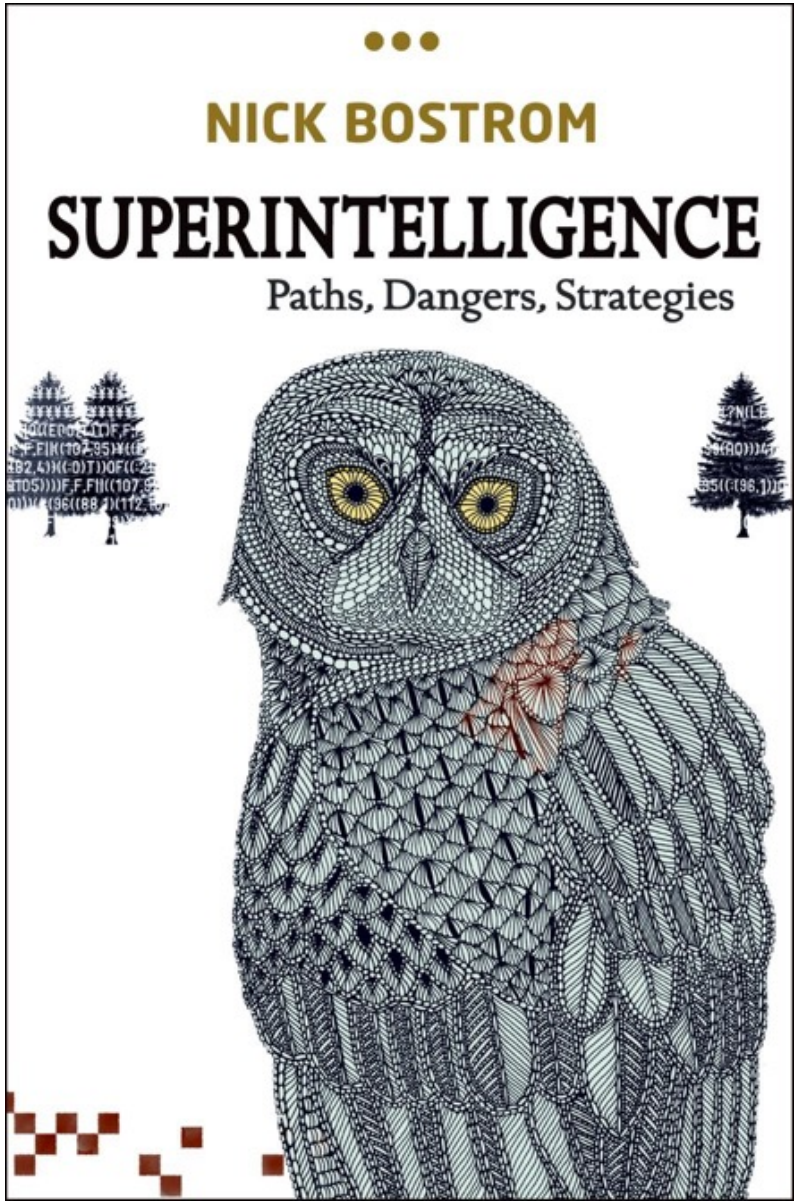
Where is the balance between the benefits and the risk associated with artificial (general) intelligence?

Trust in Digital Life Webinar

23 January 2024

Nikil Mukerji

TRUST IN DIGITAL LIFE



Amazon Customer

★☆☆☆☆ **One Star**

Reviewed in the United States on September 10, 2016

Verified Purchase

I read it in 3 days and I'm profoundly depressed.

9 people found this helpful

Helpful | Report

Agenda

1 Why Philosophy?

2 When Do We Need Philosophy?

3 Losing Control

TRUST IN DIGITAL LIFE

1 Why Philosophy?

TRUST IN DIGITAL LIFE

- Analyse concepts and questions
  - Is this concept clear?
  - Does the question we are trying to answer even make sense?
- Tease out logical implications
  - Thought experiments
- Point out logical impossibilities
- Ask (uncomfortable) questions on that basis
- Ask what *counts*
- Leave the audience profoundly depressed

1 Why Philosophy?

TRUST IN DIGITAL LIFE

When does philosophy matter?

- Usually, it does *not*
- In everyday life, we mostly are clear on
  - What our values are
  - Which questions to ask
  - What to do
- This changes when
  - Our values or our circumstances change
  - We don’t know which questions to ask
  - We are unsure what to do
- This occurs especially when the *future does not resemble the past*

2 When Do We Need Philosophy?

TRUST IN DIGITAL LIFE

We need philosophy especially when it comes to figuring out

- what *counts*
- in a *new situation*

We may call this “normative orientation”.

Example: Covid Crisis

2 When Do We Need Philosophy?

TRUST IN DIGITAL LIFE

What Counts in AGI?

- What are the central questions we need to ask?
- What are the most important aspects we should focus on?
- Where might familiar thought patterns, intuitions, heuristics etc. lead us astray?
- Which concepts and questions do we need to clarity going forward?

- If Simon is correct, the main problem is
  - *Losing Control*
  - There is no way to ensure we won't lose control.
  - The very notion of aligning AGI with our values to ensures we won't lose control is *incoherent!*

2 When Do We Need Philosophy?

TRUST IN DIGITAL LIFE

way in which they provide epistemic services to humans, notably, the provision of information and the instigation of reflection. To require that these systems be aligned to humans' *original* goals and values – those that they have prior to the AGI transition – makes no sense. To require that they be aligned with humans' evolving goals and values, as these are undergoing dramatic shifts that can purposefully be shaped by the AGI systems themselves, is too weak for making sure that humans stay in control.

We should get clear on what it means

- to stay in *control*
- lose *control*

3 Losing Control

Proposal: Losing Control is a *gradual* notion