Superintelligence

Paths, Dangers, Strategies

Nick Bostrom

Nick Bostrom

- Future of Humanity Institute, Director
- Strategic Artificial Intelligence Center, Director
- Faculty of Philosophy & Oxford Martin School, Oxford University, Professor

Fable of the Dragon Tyrant

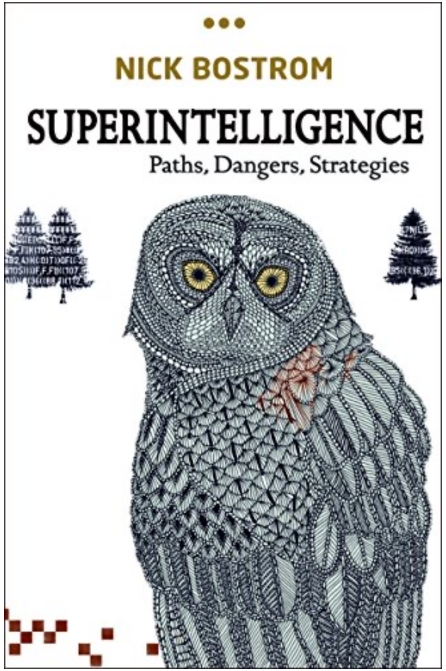
Description

General book features:

- Published in 2014
- 324 pages

Structure of the book:

- Starts with definitions and builds toward larger claims
- Each chapter addresses a different question …



Questions Adressed

- How might an entity become superintelligent?
- What are the different forms of superintelligence?
- What might an intelligence explosion look like?
- Is there a relationship between intelligence and motivation?
- Is the default outcome doom?
- How might a superintelligent entity be controlled?
- What values should superintelligent beings possess, and how might they acquire such values?
- What actions can (and should) be taken today?

Big Ideas

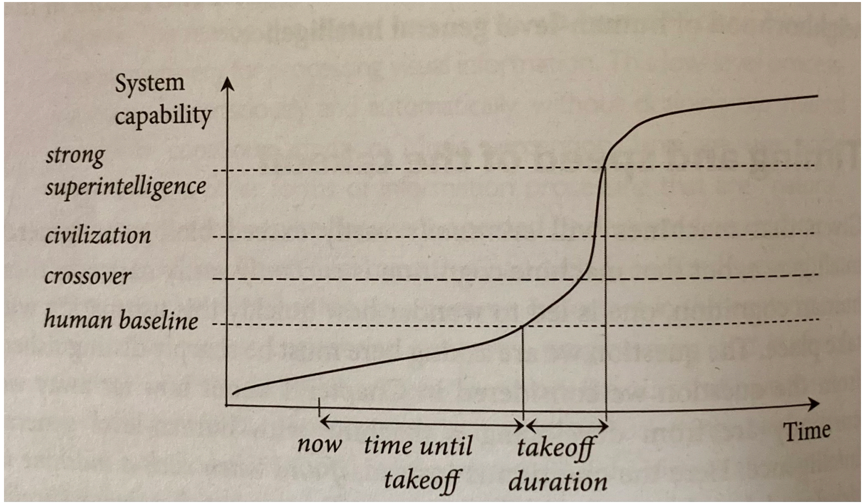
- Superintelligence will not be science fiction for long.

When will human-level machine intelligence come into existence?

study	ten_perc	fifty_perc	ninty_perc
PT-AI	2023	2048	2080
AGI	2022	2040	2065
EETN	2020	2050	2093
TOP100	2024	2050	2070
Combined	2022	2040	2075

Big Ideas (cont.)

- Without serious thought, superintelligence can be incredibly dangerous



Big Ideas (cont.)

- Don’t anthropomorphize superintelligence.
  - Capabilities
  - Motivations

MONSTER ESCAPES!
CITY FLEES IN TERROR!
ALL NEW
THRILLS! SHOCK! SUSPENSE!
REVENGE OF THE CREATURE
Universal International
starring
JOHN AGAR · LORI NELSON · JOHN BROMFIELD
with
NESTOR PAIVA · Directed by JACK ARNOLD · Screenplay by MARTIN BERKELEY · Produced by WILLIAM ALLAND · A UNIVERSAL INTERNATIONAL PICTURE
Copyright 1955 by Universal Pictures Co., Inc.
Country of Origin U.S.A.
55/122
47280

Big Ideas (cont.)

- **Orthogonality Thesis**: “more or less any level of intelligence could in principle be combined with more or less any final goal”
- **Instrumental Convergence Thesis**: “several instrumental values can be identified which are convergent in the sense that their attainment would increase the chances of the agent’s goal being realized for a wide range of final goals and a wide range of situations.”

Connection to Class

- Virtue ethics is a promising approach to creating a good ethical framework for superintelligent beings (connection to class).
  - How could an AI act virtuously?
  - Learning to live in harmony with superintelligent beings
  - Ethics as it connects to our daily and future lives, not just trollies

Recommendations

- Interesting, but not a beach read.
  - Less about story-telling, more about creating a foundation of terminology and theories
- Highly theoretical with many hypothetical situations (e.g., a superintelligent agent with the goal of creating as many paperclips as possible)
  - Great for “thinkers”, not so much for “doers”
- If you are the type of person who like thinking about what society might look like in 50 years technologically, this book is for you.