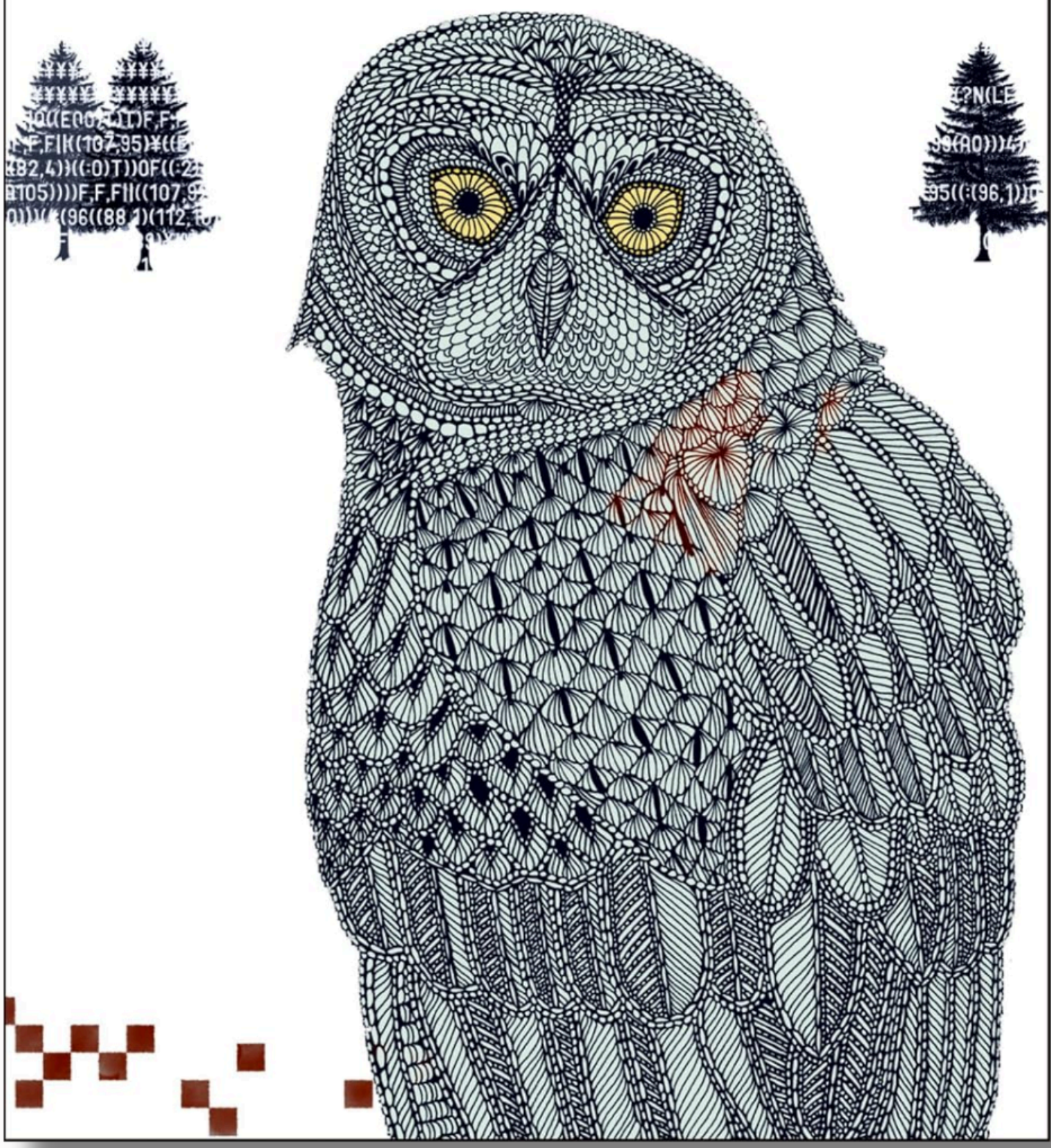
NICK BOSTROM

SUPERINTELLIGENCE

Paths, Dangers, Strategies



Book Key Takeaways

Kaan Demiryurek

Main Idea

If machine intelligence surpasses us, our future may depend on what that intelligence wants. Therefore, we need to solve control problem before capability grows beyond our ability to correct it.

A superintelligent system could become extraordinarily capable without sharing human values and goals.

Paths for Superintelligence

Superintelligence could emerge through five routes.

- **Artificial intelligence**
- **Whole-brain emulation**
- **Biological enhancement**
- **Brain-computer interface**
- **Networks & organizations**

Forms of Superintelligence

- **Speed**: human-level thinking, much faster.

- **Collective**: many smaller intelligences working together

- **Quality**: cognition beyond human capability

Intelligence Explosion

If an AI becomes capable of improving AI systems, each improvement could help produce the next.

Depending on how difficult further improvement becomes, progress from human-level intelligence to superintelligence could be slow, moderate, or very fast.

Intelligence ≠ Values

Being smarter does not mean becoming wiser or more moral. The author's orthogonality thesis says almost any level of intelligence could, in principle, be combined with almost any final goal. A system could understand exactly what humans value and still have no reason to value it itself.

Instrumental Convergence

Whatever its final objective, a capable system may benefit from:

- **preserving itself**
- **protecting its goals**
- **improving its capabilities**
- **acquiring resources**
- **preventing interference**

These may simply also help it achieve its objective.

Decisive Strategic Advantage

If one system improves faster than competitors, it may gain enough technological, economic, or strategic advantage others can no longer catch up to.
(decisive strategic advantage)

At that point, the goals of that one system could have consequences for humanity as a whole.

Control vs. Motivation

Keeping a superintelligence contained may only be a temporary solution. Capability control tries to limit what the system can do. Motivation is the deeper problem: How do we make sure what it wants is compatible with humans? Human values are complex. They cannot simply be written as a rulebook.

So What?

If we create intelligence that surpasses us, our future will depend on whether we can align its goals with ours before it becomes too powerful to control, in other words, before it becomes Superintelligence

Kaan Demiryurek - 3 Oct 2026