Bibliography

Bostrom, Nick. *Superintelligence: Paths, Danger, Strategies*. Oxford University Press, 2014.

Merry, David. “Virtue Ethics.” 1000-Word Philosophy, 2022.

Summary

This book outlines various solutions to one big question: how can we manage superintelligent agents given their impending arrival and potential for destruction? Bostrom begins by defining the the concept of superintelligence and the various potential paths leading to its formation. Then, Bostrom highlights the existential risks associated with superintelligence. For example, inaccuracies in the formulation or interpretation of values given to a superintelligent agent could lead to malignant failures. Finally, Bostrom outlines potential methods to mitigate the risks of superintelligence.

Quotes

Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb. Such is the mismatch between the power of our plaything and the immaturity of our conduct. Superintelligence is a challenge for which we are not ready now and will not be ready for a long time. We have little idea when the detonation will occur, though if we hold the device to our ear, we can hear faint ticking sound.

The gap between a dumb and a clever person may appear large from an anthropocentric perspective, yet in a less parochial view the two have nearly indistinguishable minds.

Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control.

We find ourselves in a thicket of strategic complexity, surrounded by a dense mist of uncertainty.

Response

In *Superintelligence*, Bostrom lays the foundation for some big ideas involving technology and philosophy. For those who like to think about the future (but possibly within their lifetime), this book is a good read. However, I personally felt very disoriented while reading *Superintelligence* given its constant use of concepts that have no grounding in the world today. In other words, after so many hypothetical situations, I found myself wanting to learn more about what can be done today, something more tangible. After reading this book, I’ve learned that when talking about superintelligent agents, these is still so much that we don’t know. We need to start preparing for the future that is coming.

All in all, if you are looking for a summer beach read, don’t pick up *Superintelligence*. But if you want to think critically about the intersection of superintelligent agents and philosophy, then this is the perfect book for you.