The intelligence in most of its dimensions,

Business, Decision Making

The implications of technological innovation for sustainability are becoming increasinglycomplex with information technology moving machines from being mere tools for productionor objects of consumption to playing a role in economic decision making. This emerging role willacquire overwhelming importance if artificial intelligence is underway to outperform humanintelligence in most of its dimensions, thus becoming superintelligence. Superintelligence explores the future of artificial intelligence and related technologies and therisks they may pose to human civilization. The author addresses the topic of superintelligence inthe field of artificial intelligence in both a very captivating structure and an interestingapproach. Absent the optimal background, the most important arguments of the book, will stillbe accessible, though some of the technical vocabulary may be a source of frustration for some. The book skillfully demonstrates the potential for serious thinking aimed at the long-termfuture.

A very interesting thought that the author discussed is superintelligence is compatiblewith almost any final goal. This thought is also scary somehow. This was introduced in the bookas the orthogonality thesis and has a crucial consequence that the possession ofsuperintelligence does not imply being wise or benevolent. In the matter of fact, there is noreason to believe that a superintelligent being has high moral standards; Even a non-malevolentsuperintelligence may destroy everything dear to us or perform otherwise morally terribleactions.

So if we want the self-preservation of humankind, is it enough to be careful andcautious with artificial intelligence or won’t that be enough? Is it already too late and did wealready reach the point of no return? Simulating what

human people wish requires simulatinghuman people, according to the book, terminating a simulation could then easily turn out as agenocide. For example, charged with the goal of computing the numeric value of pi, ormanufacturing many paperclips, a superintelligent being might proceed by an unbridledacquisition of physical resources, in order to facilitate its computations or manufacturingcapacity, thereby appropriating our bodies as a convenient source of atoms, or modifying ourenvironment in a way that results in our extinction. This leads to the question: what if reachingour artificial intelligence goals, whatever these may be, for comfort reasons or science reasons, will our ancestors survival goal disappear. Another theory mentioned in the book assume that superintelligences tend to converge to thesame goals. One of their reasons is that may be because most humans have similar values, alsothey will eventually adapt these values to satisfy humans. This concept is known asinstrumental convergence. Although superintelligent beings are offered a wide variety of finalgoals, they would most of the time pursue the same intermediate goal, whose satisfactiontends to enable the satisfaction of almost any final goal. Is assuming that there is a rationallycorrect morality is just us being optimistic about the future of artificial intelligence or is itrealistic that a sufficiently rational artificial intelligence will acquire this proper morality andbegin to act according to it? As much as we all want or choose to believe that humans have similar virtues, believe in samecommon good deeds and work for similar goals, just by looking at how messed up our world was and is right now (at least the way I see it based on the wars humankind witness and

theenvironmental destruction of the planet), it is hard to assume that there always exists anintermediate goal.

A goal that is good for one person, could be disastrous for another. So howcan we assume that we can make an efficient real-world artificial intelligence algorithm that issuited for the sake of everyone. Even if a human society were highly motivated to design anefficient real-world algorithm, and were given enough time to do so along with huge amountsof resources, training and knowledge about artificial intelligence, I doubt that it will come apoint where everyone could agree upon the same solution for a certain problem. How can wethen assume that artificial intelligence systems- if faced with a real problem- will take the rightdecision that will cause no harm to humans, if humans themselves fail to do this sometimes? So do I believe that superintelligence is realistic? Well, half a century after the first electriccomputer, we still have nothing that even resembles an intelligent machine, if by ‘ intelligent’we mean possessing the kind of general-purpose smartness that we humans pride ourselves on. However, neither the fact that machine intelligence would be challenging nor the fact thatsome past predictions were wrong is a good ground for concluding that artificial intelligencewill never be created.

Indeed, to assume that artificial intelligence will take thousands of yearsto develop seems at least as unwarranted as to make the opposite assumption. Already on the first few pages I sensed the speculative narrative and the intention of the authorto hold forth to conclusions such as mankind’s

doomsday. Unfortunately, there were manyspeculations that I found difficult to recognize without turning my basic logic circuits off.

Despite the book´s clear explanation of why superintelligent AI may have arbitrarily negativeconsequences and why it is important to begin addressing the issue well in advance, the authordoes not base his case on predictions that superhuman AI systems are imminent. He writes, “ Itis no part of the argument in this book that we are on the threshold of a big breakthrough inartificial intelligence, or that we can predict with any precision when such a development mightoccur.” Pointing to long-term risks from AI is however not equivalent to claiming thatsuperintelligence and its accompanying risks are imminent. From my point of you view, I do think that AI presents an existential risk and it should beseriously recognized. This risk does not necessarily rise from spontaneous maliciousconsciousness, but rather from the unpredictability and potential irreversibility of deploying anoptimization process more intelligent than the humans who specified its objectives. However, no one can say for sure if it is 100% happening and when will it happen. We all baseour opinions on assumptions depending on how optimistic or not we are and according to ourown knowledge of the evolution of technological inventions. But the future is alwaysunpredictable.

One thing is for sure the challenge of superintelligence cannot be separatedfrom the other major environmental and social challenges, demanding a fundamentaltransformation along the lines of degrowth. With machines outperforming humans in theirfunctions, maybe our social skills will still keep humans ahead of machines. As a conclusion, my general

impression was that this book might not be the most pleasurableread at some sections, sometimes repetitive and not straightforward, nevertheless, stickingwith it till the end was for me worth its while. The book has a good enough logical structure, and most chapters end with a summary.

The author successed in giving a detailed analysisabout this very important theme of the modern world and engaged with the topic verycompetently. To my viewpoint Bostrom succeeds in arguing that the development ofsuperintelligent machines will, if not properly managed, create catastrophic risks to humanity.