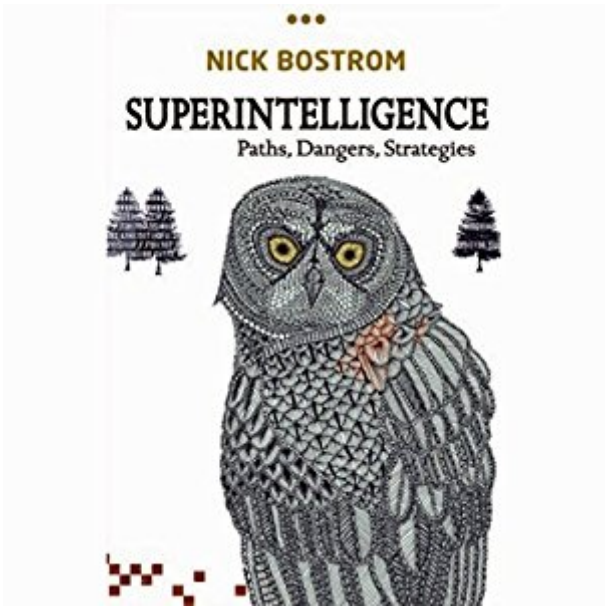
Superintelligence: Paths, Dangers, Strategies



Synopsis

Superintelligence asks the questions: What happens when machines surpass humans in general intelligence? Will artificial agents save or destroy us? Nick Bostrom lays the foundation for understanding the future of humanity and intelligent life. The human brain has some capabilities that the brains of other animals lack. It is to these distinctive capabilities that our species owes its dominant position. If machine brains surpassed human brains in general intelligence, then this new superintelligence could become extremely powerful - possibly beyond our control. As the fate of the gorillas now depends more on humans than on the species itself, so would the fate of humankind depend on the actions of the machine superintelligence. But we have one advantage: We get to make the first move. Will it be possible to construct a seed Artificial Intelligence, to engineer initial conditions so as to make an intelligence explosion survivable? How could one achieve a controlled detonation? This profoundly ambitious and original book breaks down a vast track of difficult intellectual terrain. After an utterly engrossing journey that takes us to the frontiers of thinking about the human condition and the future of intelligent life, we find in Nick Bostrom’s work nothing less than a reconceptualization of the essential task of our time.

Book Information

Audible Audio Edition
Listening Length: 14 hoursÂ andÂ 16 minutes
Program Type: Audiobook
Version: Unabridged
Publisher: Audible Studios
Audible.com Release Date: September 3, 2014
Language: English
ASIN: B00LPMFE9Y
Best Sellers Rank: #4 inÂ Books > Computers & Technology > Computer Science > AI & Machine Learning > Intelligence & Semantics #9 inÂ Books > Audible Audiobooks > Nonfiction > Computers

Customer Reviews

Prof. Bostrom has written a book that I believe will become a classic within that subarea of Artificial Intelligence (AI) concerned with the existential dangers that could threaten humanity as the result of the development of artificial forms of intelligence.What fascinated me is that Bostrom has approached the existential danger of AI from a perspective that, although I am an AI professor, I had never really examined in any detail.When I was a graduate student in the early 80s, studying for my

PhD in AI, I came upon comments made in the 1960s (by AI leaders such as Marvin Minsky and John McCarthy) in which they mused that, if an artificially intelligent entity could improve its own design, then that improved version could generate an even better design, and so on, resulting in a kind of "chain-reaction explosion" of ever-increasing intelligence, until this entity would have achieved "superintelligence". This chain-reaction problem is the one that Bostrom focusses on. He sees three main paths to superintelligence:1. The AI path -- In this path, all current (and future) AI technologies, such as machine learning, Bayesian networks, artificial neural networks, evolutionary programming, etc. are applied to bring about a superintelligence.2. The Whole Brain Emulation path -- Imagine that you are near death. You agree to have your brain frozen and then cut into millions of thin slices. Banks of computer-controlled lasers are then used to reconstruct your connectome (i.e., how each neuron is linked to other neurons, along with the microscopic structure of each neuron's synapses). This data structure (of neural connectivity) is then downloaded onto a computer that controls a synthetic body. If your memories, thoughts and capabilities arise from the connectivity structure and patterns/timings of neural firings of your brain, then your consciousness should awaken in that synthetic body.The beauty of this approach is that humanity would not have to understand how the brain works. It would simply have to copy the structure of a given brain (to a sufficient level of molecular fidelity and precision).3. The Neuromorphic path -- In this case, neural network modeling and brain emulation techniques would be combined with AI technologies to produce a hybrid form of artificial intelligence. For example, instead of copying a particular person's brain with high fidelity, broad segments of humanity's overall connectome structure might be copied and then combined with other AI technologies.Although Bostrom's writing style is quite dense and dry, the book covers a wealth of issues concerning these 3 paths, with a major focus on the control problem. The control problem is the following: How can a population of humans (each whose intelligence is vastly inferior to that of the superintelligent entity) maintain control over that entity? When comparing our intelligence to that of a superintelligent entity, it will be (analogously) as though a bunch of, say, dung beetles are trying to maintain control over the human (or humans) that they have just created.Bostrom makes many interesting points throughout his book. For example, he points out that a superintelligence might very easily destroy humanity even when the primary goal of that superintelligence is to achieve what appears to be a completely innocuous goal. He points out that a superintelligence would very likely become an expert at dissembling -- and thus able to fool its human creators into thinking that there is nothing to worry about (when there really is).I find Bostrom's approach refreshing because I believe that many AI researchers have been either unconcerned with the threat of AI or they have focussed only on the threat to humanity once a large

population of robots is pervasive throughout human society.I have taught Artificial Intelligence at UCLA since the mid-80s (with a focus on how to enable machines to learn and comprehend human language). In my graduate classes I cover statistical, symbolic, machine learning, neural and evolutionary technologies for achieving human-level semantic processing within that subfield of AI referred to as Natural Language Processing (NLP). (Note that human "natural" languages are very very different from artificially created technical languages, such a mathematical, logical or computer programming languages.)Over the years I have been concerned with the dangers posed by "run-away AI" but my colleagues, for the most part, seemed largely unconcerned. For example, consider a major introductory text in AI by Stuart Russell and Peter Norvig, titled: Artificial Intelligence: A Modern Approach (3rd ed), 2010. In the very last section of that book Norvig and Russell briefly mention that AI could threaten human survival; however, they conclude: "But, so far, AI seems to fit in with other revolutionary technologies (printing, plumbing, air travel, telephone) whose negative repercussions are outweighed by their positive aspects" (p. 1052).In contrast, my own view has been that artificially intelligent, synthetic entities will come to dominate and replace humans, probably within 2 to 3 centuries (or less). I imagine three (non-exclusive) scenarios in which autonomous, self-replicating AI entities could arise and threaten their human creators.(1) The Robotic Space-Travel scenario: In this scenario, autonomous robots are developed for space travel and asteroid mining. Unfortunately, many people believe in the alternative "Star Trek" scenario, which assumes that: (a) faster-than-light (warp drive) will be developed and (b) the galaxy will be teeming, not only with planets exactly like Earth, but also these planets will be lacking any type of microscopic life-forms dangerous to humans. In the Star Trek scenario, humans are very successful space travelers.However, It is much more likely that, to make it to a nearby planet, say, 100 light years away, will require that humans travel for a 1000 years (at 1/10th the speed of light) in a large metal container, all the while trying to maintain a civilized society as they are being constantly radiated while they move about within a weak gravitational field (so their bones waste away while they constantly recycle and drink their urine). When their distant descendants finally arrive at the target planet, these descendants will very likely discover that the target planet is teeming with deadly, microscopic parasites.Humans have evolved on the surface of the Earth and thus their major source of energy is oxygen. To survive they must carry their environment around with them. In contrast, synthetic entities will require no oxygen or gravity. They will not be alive (in the biological sense) and so therefore will not have to expend any energy during the voyage. A simple clock can turn them on once they have arrived at the target planet and they will be unaffected by any forms of alien microbial life.If there were ever a conflict between humans and these space-traveling synthetic

AI entities, who would have the advantage? The synthetic entities would be looking down on us from outer space -- a definitive advantage. (If an intelligent alien ever visits Earth, it is 99.9999% likely that whatever exits the alien spacecraft will be a non-biological, synthetic entity -- mainly because space travel is just too difficult for biological creatures.)(2) The Robotic Warfare scenario: No one wants their (human) soldiers to die on the battlefield. A population of intelligent robots that are designed to kill humans will solve this problem. Unfortunately, if control over such warrior robots is ever lost, then this could spell disaster for humanity.(3) The Increased Dependency scenario: Even if we wanted to, it is already impossible to eliminate computers because we are so dependent on them. Without computers our financial, transportation, communication and manufacturing services would grind to a halt. Imagine a near-future society in which robots perform most of the services now performed by humans and in which the design and manufacture of robots are handled also by robots. Assume that, at some point, a new design results in robots that no longer obey their human masters. The humans decide to shut off power to the robotic factory but it turns out that the hydroelectric plant (that supplies it with power) is run by robots made at that same factory. So now the humans decide to halt all trucks that deliver materials to the factory, but it turns out that those trucks are driven by robots, and so on.I had always thought that, for AI technology to pose an existential danger to humanity, it would require processes of robotic self-replication. In the Star Trek series, the robot Data is more intelligence that many of his human colleagues, but he has no desire to make millions of copies of himself, and therefore he poses less of a threat than, say, south american killer bees (which have been unstoppable as they have spread northward).Once synthetic entities have a desire to improve their own designs and to reproduce themselves, then they will have many advantages over humans: Here are just a few:1. Factory-style replication: Humans require approximately 20 years to produce a functioning adult human. In contrast, a robotic factory could generate hundreds of robots every day. The closest event to human-style (biological) replication will occur each time a subset of those robots travel to a new location to set up a new robotic factory.2. Instantaneous learning: Humans have always dreamt of a "learning pill" but, instead, they have to undergo that time-consuming process called "education". Imagine if one could learn how to fly a plane just by swallowing a pill. Synthetic entities would have this capability. The brains of synthetic entities will consist of software that executes on universal computer hardware. As a result, each robot will be able to download additional software/data to instantly obtain new knowledge and capabilities.3. Telepathic communication: Two robots will be able communicate by radio waves, with robot R1 directly transmitting some capability (e.g., data and/or algorithms learned through experience) to another robot R2.4. Immortality: A robot could back up a copy of its mind

(onto some storage device) every week. If the robot were destroyed, a new version could be reconstructed with just the loss of one week’s worth of memory.5. Harsh Environments: Humans have developed clothing in order to be able to survive in cold environments. We go into a closet and select thermal leggings, gloves, goggles, etc. to go snowboarding. In contrast, a synthetic entity could go into its closet and select an alternative, entire synthetic body (for survival on different planets with different gravitational fields and atmospheres).What is fascinating about Bostrom’s book is that he does not emphasize any of the above. Instead, he focusses his book on the dangers, not from a society of robots more capable than humans, but, instead, on the dangers posed by a single entity with superintelligence coming about. (He does consider what he calls the "multipolar" scenario, but that is just the case of a small number of competing superintelligent entities.)Bostrom is a professor of philosophy at Oxford University and so the reader is also treated to issues in morality, economics, utility theory, politics, value learning and more.I have always been pessimistic about humanity’s chance of avoiding destruction at the hands of it future AI creations and Bostrom’s book focusses on the many challenges that humanity may (soon) be facing as the development of a superintelligence becomes more and more likely.However, I would like to point out one issue that I think Prof. Bostrom mostly overlooks. The issue is Natural Language Processing (NLP). He allocates only two sentences to NLP in his entire book. His mention of natural language occurs in Chapter 13, in his section on "Morality models". Here he considers that, when giving descriptions to the superintelligence (of how we want it to behave), its ability to understand and carry out these descriptions may require that it comprehend human language, for example, the term "morally right".He states:"The path to endowing an AI with any of these concepts might involve giving it general linguistic ability (comparable, at least, to that of a normal human adult). Such a general ability to understand natural language could then be used to understand what is meant by ’morally right’ " (p. 218)I fear that Bostrom has not sufficiently appreciated the requirements of natural language comprehension and generation for achieving general machine intelligence. I don’t believe that an AI entity will pose an existential threat until it has achieved at least a human level of natural language processing (NLP).Human-level consciousness is different than animal-level consciousness because humans are self-aware. They not only think thoughts about the world; they also think thoughts about the fact that they are thinking thoughts. They not only use specific words; they are aware of the fact that they are using words and how different categories of words differ in functionality. They are not only capable of following rules; they are aware of the fact that rules exist and that they are able to follow (or not follow) those rules. Humans are able to invent and modify rules.Language is required to achieve this level of self-reflective thought and creativity. I define

(human-level natural) language as any system in which the internal structures of thought (whatever those happen to be, whether probabilities or vectorial patterns or logic/rule structures or dynamical attractors or neural firing patterns, etc.) are mapped onto external structures -- ones that can then be conveyed to others.Self-awareness arises because this mapping enables the existence of a dual system:Internal (Thought) Structures External (Language) Structures.In the case of human language, these external structures are symbolic. This dual system enables an intelligent entity to take the results of its thought processes, map them to symbols and then use these symbols to trigger thoughts in other intelligent entities (or in oneself). An entity with human-level self-awareness can hold a kind of conversation with itself, in which it can refer to and thus think about its own thinking.Something like NLP must therefore exist BEFORE machines can reach a level of self-awareness to pose a threat to humanity. In the case of a super-intelligence, this dual system may look different than human language. For example, a superintelligence might map internal thoughts, not only to symbols of language, but also to complex vectorial structures. But the point is the same -- something must act like an external, self-referential system -- a system than can externally refer to the thoughts and processes of that system itself.In the case of humans, we do not have access to the internal structure of our own thoughts. But that doesn't matter. What matters is that we can map aspects of our thoughts out to external, symbolic structures. We can then communicate these structures to others (and also back to ourselves). Words/sentences of language can then trigger thoughts about the world, about ourselves, about our goals, our plans, our capabilities, about conflicts with others, about potential future events, about past events, etc.Bostrom seems to imply (by his oversight) that human-level (and super-human levels) of general intelligence can arise without language. I think this is highly unlikely.An AI system with NLP capability makes the control problem much more difficult than even Bostrom claims. Consider a human H1 who kills others because he believes that God has commanded him to kill those with different beliefs. Since he has human-level self-awareness, he should be explicitly aware of his own beliefs. If H1 is sufficiently intelligent then we should be able to communicate a counterfactual to H1 of the sort: "If you did not believe in God or if you did not believe that God commanded you to kill infidels, then you would not kill them." That is, H1 should have access (via language) to his own beliefs and to knowledge into how changes in those beliefs might (hypothetically) change his own behavior.It is this language capability that enables a person to change their own beliefs (and goals, and plans) over time. It is the combination of the self-reflective nature of human language, combined with human learning abilities, that makes it extremely difficulty to both predict and control what humans will end up believing and/or desiring (let alone superintelligent entities)It is extremely

difficult but (hopefully) not impossible to control a self-aware entity. Consider two types of psychiatric patients: P1 and P2. Both have a compulsion to wash their hands continuously. P1 has what doctors call "insight" into his own condition. P1 states: "I know I am suffering from an obsessive/compulsive trait. I don't want to keep washing my hands but I can't help myself and I am hoping that you, the doctors, will cure me." In contrast, patient P2 lacks "insight" and states: "I'm fine. I wash my hands all the time because it's the only way to make be sure that they are not covered with germs."If we were asked which patient appears more intelligent (all other things being equal) we would choose P1 as being more intelligent than P2 because P1 is aware of features of P1's own thinking processes (that P2 is not aware of).As a superintelligent entity becomes more and more superintelligent, it will have more and more awareness of its own mental processes. With increased self-reflection it will become more and more autonomous and less able to be controlled. LIke humans, it will have to be persuaded to believe in something (or to take a certain course of action). Also, this superintelligent entity will be designing even more self-aware versions of itself. Increased intelligence and increased self-reflection go hand in hand. Monkeys don't persuade humans because monkeys lack the ability to refer to the concepts that humans are able to entertain. To a superintelligent entity we will be as persuasive as monkeys (and probably much less persuasive) .Any superintelligent entity that incorporates human general intelligence will exhibit what is commonly referred to as "free will". Personally, I do not believe that my choices are made "freely". That is, my neurons fire -- not because they choose to, but because they had to (due to the laws of physics and biochemistry). But let us define "free will" as any deterministic system with the following components/capabilities:a. The NLP ability to understand and generate words/sentences that refer to its own thoughts and thought processes, e.g. to be able to discuss the meaning of the word "choose".b. Ability to generate hypothetical, possible futures before taking an action and also, ability to generate hypothetical, alternative pasts after having taken that action.c. Ability to think/express counterfactual thoughts, such as "Even though I chose action AC1, I could have instead chosen AC2, and if I had done so, then the following alternative future (XYZ) would likely have occurred."Such as system (although each component is deterministic and so does not violate the laws of physics) will subjectively experience having "free will". I believe that a superintelligence will have this kind of "free will" -- in spades.Given all the recent advances in AI (e.g. autonomous vehicles, object recognition learning by deep neural networks, world master-level play at the game of Jeopardy by the Watson program, etc.) I think that Bostrom's book is very timely.Michael Dyer

Not surprisingly, 200+ pages later, the author can't answer the 'what is to be done' question

concerning the likely emergence of non-human (machine-based) super-intelligence, sometime, possibly soon. This is expected because, as a species, we've always been the smartest ones around and never had to even think about the possibility of coexistence alongside something or someone impossibly smart and smart in ways well beyond our comprehension, possibly driven by goals we can't understand and acting in ways that may cause our extinction.Building his arguments on available data and extrapolating from there, Bostrom is confident that:- some form of self-aware, machine super-intelligence is likely to emerge- we may be unable to stop it, even if we wanted to, no matter how hard we tried- while we may be unable to stop the emergence of super-intelligence, we could prepare ourselves to manage it and possibly survive it- us not taking this seriously and not being prepared may result in our extinction while serious pre-emergence debate and preparation may result in some form of co-existenceIt's radical and perhaps frightening but our failure to comprehend the magnitude of the risks we are about to confront would be a grave error given that, once super-intelligence begins to manifest itself and act, the change may be extremely quick and we may not be afforded a second chance.Most of the book concerns itself with the several types of super-intelligence that may develop, the ways in which we may be able to control or at least co-exist with such entities or entity, what the world and literally the Universe may turn into depending on how we plant the initial super-intelligent seed. The author also suggests that it may be possible for us to survive and even benefit if we manage to do everything just about right. Of course, the odds of that happening given human nature are extremely small but some optimism is needed or we'd just give up and allow ourselves to go extinct or perhaps all turn into maintenance workers, serving our all-knowing, all-powerful master.I am not going to go into any further detail. Bostrom makes his case with competence and humor and this well-researched, original and important work deserves to be read and understood and, hopefully, taken seriously enough for some of us to expand upon his research and act upon our conclusions. I will end my little review here but not before I quote from Bostrom's book, his eloquent warning: "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"How did he come to such a radical and pessimistic conclusion? You better read the book. It's not exactly fifth grade level material but it can be a fascinating read for anyone sufficiently motivated, patient and open-minded.

Hidden Dangers of Cosmetics Summary: The Plant Paradox: The Hidden Dangers in "Healthy" Foods That Cause Disease and Weight Gain by Steven R. Gundry M.D. Less Pain, Fewer Pills: Avoid the Dangers of Prescription Opioids and Gain Control over Chronic Pain Overcoming Lyme Disease: The Truth About Lyme Disease and The Hidden Dangers Plaguing Our Bodies Rheumatoid Arthritis Unmasked: 10 Dangers of Rheumatoid Disease A Narrative of a Revolutionary Soldier: Some Adventures, Dangers, and Sufferings of Joseph Plumb Martin (Signet Classics) Toxic Mold: Beware Of The Dangers Of Mold All God's Dangers The Gray Rhino: How to Recognize and Act on the Obvious Dangers We Ignore Summary, Analysis, and Review of Steven R. Gundry's The Plant Paradox: The Hidden Dangers in "Healthy" Foods That Cause Disease and Weight Gain Exploring Dangers in Space: Asteroids, Space Junk, and More The Unhealthy Truth: One Mother's Shocking Investigation into the Dangers of America's Food Supply-- and What Every Family Can Do to Protect Itself Not So Fast: Parenting Your Teen Through the Dangers of Driving Punch: The Delights (and Dangers) of the Flowing Bowl How to Talk to Your Cat About Gun Safety: And Abstinence, Drugs, Satanism, and Other Dangers That Threaten Their Nine Lives Global Problems and Dangers Domestic Dangers: Women, Words, and Sex in Early Modern London (Oxford Studies in Social History) Shareveillance: The Dangers of Openly Sharing and Covertly Collecting Data (Forerunners: Ideas First)