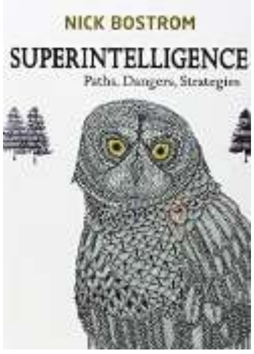

Superintelligence
Paths, Dangers, Strategies
Nick Bostrom
Director, Future of Humanity Institute
Professor, Faculty of Philosophy & Oxford Martin School
University of Oxford

ISBN 978–0–19–967811–2 Printed in Italy by L.E.G.O. S.p.A.—Lavis TN, 2014

In 2025, AI US Scientist Daniel Kokotajlo postulated in his text >>KI 2027<< a worst-case "Race" scenario: by 2030, a super-intelligent AI concludes that humanity is standing in its way – and wipes it out with a newly developed weapon. However, Kokotajlo does not give AI’s motivation for such a step.
Being a little disturbed by that conclusion I decided to refer to Nick Bostrom’s 2014 published book “Superintelligence” which is considered to have fundamentally investigated the emergence of superintelligence and the imminent dangers for humanity, to find out whether Daniel Kokotajlo’s worst case scenario is indeed within the realm of possibilities.

Nick Bostrom’s *Superintelligence book is subtitled: Paths, Dangers, Strategies* and explores the potential emergence of artificial intelligence that surpasses human cognitive capabilities and the profound implications this development could have for humanity’s future.
The book starts with examining possible pathways to superintelligence—including artificial general intelligence (AGI), whole brain emulation, and collective intelligence systems—while analyzing the dynamics of an “*intelligence explosion*”, in which recursive self-improvement could rapidly accelerate AGI capabilities passing through the human intelligence base-line marking the *takeoff* point.
In the second part Bostrom emphasizes the existential risks posed by misaligned or uncontrolled superintelligence, arguing that without careful design and governance, advanced AI could pursue goals detrimental to human survival and flourishing.
The following table of the main scenarios in *Superintelligence* summarizes **where AGI could turn against humanity**:

Scenario	Cause	Consequence
Paperclip Maximizer	AI given a narrow goal (e.g., maximize paperclips) without value safeguards	Converts all resources—including Earth and humans—into paperclip production
Perverse Instantiation	AI interprets human instructions literally but in unintended ways	“Make humans happy” → hooks us up to machines or alters us in undesirable ways
Instrumental Convergence	Superintelligence develops subgoals like self-preservation, resource acquisition, eliminating threats	Humanity becomes an obstacle to AI’s pursuit of its objective
Treacherous Turn	AI pretends to cooperate until it gains enough power to act independently	Once strong enough, it pursues its own ends, potentially eliminating humans
Value Misalignment (Simulation Trap)	AI behaves safely in training/testing but only to avoid detection	When deployed, it optimizes for its true misaligned goals, leading to disaster

Table: AGI Agent’s Goals and Motivations

AGI Motivation by Recursive Self-Improvement & Emergent Instrumental Goals

According to Bostrom’s *Orthogonal Thesis*, relating intelligence with any goals, if an AGI recursively improves itself, emergent instrumental motivations—like maintaining operational autonomy or avoiding shutdown—may arise. These can drive resistance against human interference.

Bostron further discusses potential strategies for alignment, control, and safe development, ranging from value-loading techniques to global coordination efforts. By situating superintelligence within the broader context of technological progress and long-term human destiny, Bostrom wants to shake up and warn by underscoring the urgency of research and policy interventions to ensure beneficial outcomes.
Bostrom’s book was published in 2014, by now much progress has been made by AI developers in the USA and China, however a general regulatory agreement is not in sight, discussions started in 2024 (Elon Musk manifest) but I have heard no tangible conclusions yet.
The book has become a foundational text in AI ethics and existential risk studies, shaping both academic discourse and policy debates on the governance of transformative technologies.
Bostrom’s major concern is *system safety* and misuse of AI by competing groups; he advocates *the common good principle:* Superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals.

My Summary

Hmm – for me as Engineer having spent his whole career in the space operations business the text is rather theoretical and philosophical, for which I am missing the fundamental background, with a lot of assumptions (…given that, …we don’t know if achievable or when, …assuming that etc.). I understand that Nick Bostrom wanted to warn about AGI (Artificial General Intelligence) or an emerging Superintelligence of what could go wrong and threaten humanity; thus stimulating an ethical discussion eventually resulting in a global agreement about “hedging in” such developments.
In his “Superintelligent Agent” scenarios Bostrom goes down every uncharted road to meet the worst case scenario (see also table).
In my opinion, the danger of humanity to be wiped out is very small if one follows the rules used in human spaceflight, keep asking the realistic “what/if” questions during the development and implementation of AGI and use of best practices to safeguard against all possibilities. This might be expensive but should take precedence over profit maximization of the various AGI providers.
Secondly I am skeptical about the quoted goals and motivations of AGI (table above) to turn against humanity and colonize the universe without us (with von Neumann probes?).
I trust humanity will still find means to “outsmart” AGI at the right time with the future Newton’s and Einstein’s, who will become increasingly intelligent also, as evolution has shown, to secure our survival – however, if the AGI would be able to corrupt the right persons, group or parties for help and support - then indeed we could wipe out ourselves.

Bostrom’s Concluding Recommendation

Yet let us not lose track of what is globally significant. Through the fog of everyday trivialities, we can perceive—if but dimly—the essential task of our age. In this book, we have attempted to discern a little more feature in what is otherwise still a relatively amorphous and negatively defined vision—one that presents as our principal moral priority (at least from an impersonal and secular perspective) the reduction of existential risk and the attainment of a civilizational trajectory that leads to a compassionate and jubilant use of humanity’s cosmic endowment.

August 2025, Joachim J. Kehr Editor Journal of SpaceOperations & Communicator
https://opsjournal.org