Nick Bostrom's "Superintelligence": The Central Argument

The focal point of the argumentation is the thesis:

[**Superintelligence as existential risk**]: A plausible *default* outcome of a (machine) superintelligence explosion -- the latter being itself a serious possibility -- is existential catastrophe, i.e. the extinction of human civilization.

The Argumentation in favor of @[Superintelligence as existential risk]

[**Superpower**]: A seed superintelligence has the cognitive and physical means to bring any resources on earth, including humans, under its control (cf. Chapter 6).

[**Orthogonality thesis**]: Cognitive abilities are largely independent from -- and do not determine -- specific final goals an intelligent system pursues (cf. Chapter 7).

[**Instrumental convergence thesis**]: All superintelligent systems, no matter what their final goals, are likely to share certain intermediate goals, namely the acquisition of as many physical resources as possible in order to reach their final goals (cf. Chapter 7).

<**Take-over-scenario argument**>: A superintelligent system is likely to bring under control all resources on Earth, reducing humans to mere means to obtain non-anthropomorphic final goals, and thus leading to the extinction of human civilization as we know it (cf. Chapter 8).

- **+** [**Superpower**]
- **+** [**Orthogonality thesis**]
- **+** [**Instrumental convergence thesis**]
- **+>** [**Superintelligence as existential risk**]

<**Superintelligence as serious possibility**>: Given alternative technological realizations and the positive feedbacks involved in the development of intelligent systems, the creation of a superintelligent system within this century is a serious possibility (cf. Chapters 2-5).

- **+>** <**Take-over-scenario argument**>

[**Alternative paths to SI**]: There exist alternative technological paths to superintelligent systems (cf. Chapter 2).

+> <**Superintelligence as serious possibility**>

[**Bootstrapping**]: Intelligent systems can (be used to) further improve their cognitive abilities, creating positive feedbacks in SI development (cf. Chapters 4,5).

+> <**Superintelligence as serious possibility**>

@[Superintelligence as existential risk] Gives Rise to the Control Problem

Bostrom's main claim concerning the control problem, and its justification are:

[**Value-based control methods required**]: We need to develop methods for controlling superintelligent systems by designing their value system (which involves itself normative assumptions).

[**Control methods**]: Intelligent systems can be controlled either through capability containment or motivation selection (cf. Chapter 9).

+> <**Control Problem Argument**>

<**Capability control ineffective**>: Capability control is ineffective and unreliable for superintelligent systems, because the SI is likely to see ways to manipulate the world which we don't (cf. Chapters 9, 12, 13).

+> <**Control Problem Argument**>

<**Control Problem Argument**>: Motivation selection is the only control method that has a chance to avoid the doomsday take-over scenario (cf. Chapter 9).

+ [**Superintelligence as existential risk**]

+> [**Value-based control methods required**]

Science and Technology Policies

Decision-makers may entertain the following policies:

[**Slower SI development!**]: We should retard the development of machine SI.

[**Slower cognitive enhancement development!**]: We should retard the

development of cognitive enhancement.

What are the pros and cons?

A relevant general principle in this context is:

[**The principle of differential technological development**]: Retard the development of dangerous and harmful technologies, especially ones that raise the level of existential risk; and accelerate the development of beneficial technologies, especially those that reduce the existential risks posed by nature or by other technologies (cf. p. 282).

Pros and Cons Machine SI

<**Machine SI development too risky**>: The development of machine SI poses existential risks.

+> [**Slower SI development!**]
+ [**The principle of differential technological development**]
+ [**Superintelligence as existential risk**]

<**No societal preparation without R&D**>: Only once R&D efforts into SI are underway, will society start to prepare seriously for the advent of superintelligence; so the earlier R&D starts, the more time we have to get ready (cf. pp. 293-4).

(1) The risks of machine SI are great.

+ [**Superintelligence as existential risk**]

(2) Reducing these risks will require a period of serious preparation.

(3) Serious preparation will begin only once the prospect of machine SI is taken seriously by broad sectors of society.

(4) Broad sectors of society will take the prospect of SI seriously only once a large research effort to develop machine SI is underway.

(5) The earlier a serious research effort is initiated, the longer it will take to deliver machine SI.

+ Because R&D starts from a lower level of pre-existing enabling technologies.

---

(6) The earlier a serious research effort is initiated, the longer the period during which serious preparation will take place.

(7) The more time to prepare the better.

(8) **A serious reserach effort toward machine SI should be initiated.**

-> [**Slower SI development!**]

Pros and Cons Cognitive Enhancement

<**Cognitive enhancement development too risky**>: The development of cognitive enhancement poses existential risks, namely inasmuch as it represents a path to superintelligence.

+> [**Slower cognitive enhancement development!**]
\+ [**The principle of differential technological development**]
\+ [**Superintelligence as existential risk**]

<**Cognitive enhancement prepares us for SI**>: The development of cognitive enhancement reduces an existential risk, namely inasmuch as it prepares us for a superintelligence explosion -- specifically by enabling us to tackle the control problem (cf. pp. 288-90).

-> [**Slower cognitive enhancement development!**]
\+ [**The principle of differential technological development**]
\+ [**Superintelligence as existential risk**]
\+ [**Value-based control methods required**]