The Monster in the Library of Turing

Thore Husfeldt

April 10, 2015

To appear (in Swedish) as T. Husfeldt, *Monstret i Turings bibliotek*, Filosofisk tidskrift, nr. XX, 2015.

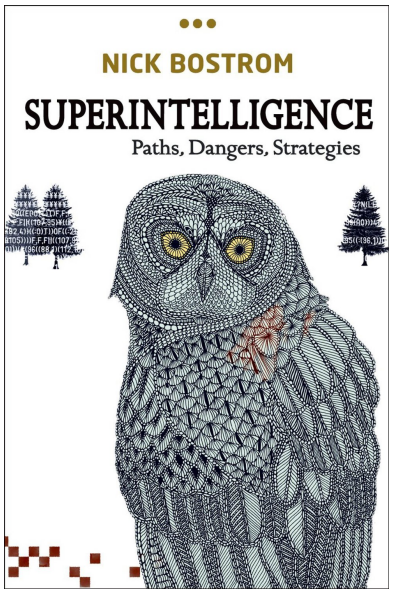

Figure 1: Nick Bostrom, *Superintelligence*, Oxford University Press, 2014.

Nick Bostrom's *Superintelligence* presents a dystopian view of strong artificial intelligence: Not only will it be the greatest invention of mankind, it will also be the last, and this emerging technology should be viewed as an immediate and catastrophic risk to our species, like molecular nanotechnology, nuclear power, or chemical warfare.

Artificial intelligence with super-human capabilities will be the last invention controlled by Homo sapiens, since all subsequent inventions will be made by the "superintelligence" itself. Moreover, unless we are extremely careful or lucky, the superintelligence will destroy us, or at least radically change our living conditions in a way that we may find undesirable. Since we are currently investigating many technologies that may lead to a superintelligence, now would be a good time for reflection.

Nobody knows, much less agrees on, how to define intelligence, be it general, artificial, or strong. Neither does Bostrom. By his own admission, his book is inherently speculative and probably wrong. Not even the rudiments of the relevant technology may be known today. However, many of Bostrom's arguments are quite robust to the particulars of "what?" and "how?", and the book can be enjoyed without a rigorous definition. For now, we imagine the superintelligence as a hypothetical agent that is much smarter than the best current human brains in every cognitive activity, including scientific creativity, general wisdom, and social skills.

Pathways and Consequences

Bostrom describes a number of pathways towards a superintelligence. They consist of current scientific activities and emerging technologies, extrapolated beyond human cognitive capabilities. These include

- neuroanatomical approaches, such as simulation of whole brains (Einstein's brain in a vat or simulated on a computer, neurone by neurone),
- genetically or artificially modified humans (brain implants, parental gamete selection for intelligence),
- intelligence as an emergent phenomenon in simpler networks (the internet as a whole "becomes conscious"),
- computer science approaches like machine learning (IBM's Jeopardy-winning *Watson*) and "good old-fashioned artificial intelligence" using symbolic reasoning (chess computers).

None of these technologies is currently close to even a dullard's intelligence, but on a historical time scale, the are all very new. Bostrom's description is an entertaining tour of computer science, neuroanatomy, evolutionary biology, cognitive psychology, and related fields. It is moderately informative, but none of it is authoritative.

The consequences of any of these research programs actually achieving their goals are even more speculative. Several futurists have developed scenarios for how a superintelligent future might look, and Bostrom surveys many of these ideas. In one of those visions, accelerating and self-improving digital technologies quickly overtake human cognitive abilities and transform our environment, like we transformed pre-historic Earth. After a short while, all the protons that currently making up the planets of the solar system are put to better use as solar-powered computing devices, orbiting the sun in a vast networked intelligence known as a Dyson sphere. There are many other scenarios, partly depending on which approach will turn out to win the race against our own intelligence, and how the result can be harnessed. Not all these possible futures are dystopian. Some even leave room for humans, maybe uploaded to a digital version of immortality, or kept in zoos. But they certainly entail dramatic changes to our way of life, easily comparable to the invention of agriculture or the industrial revolution.

The Control Problem

Bostrom' book begins with a short fable, in which a group of birds agrees to look for an owl chick to help them with nest-buildling and other menial labours. We immediately see their folly: The owl chick will quickly outgrow the birds, throw off their yoke, and quite probably eat them, as is its nature. The birds would have been better off thinking about the result of their search before it was too late.

However, in its most harmless form, the superintelligence is merely a device that excels at goal-oriented behaviour, without intentionality, consciousness, or predatory tendencies. Bostrom explains that even such a superintelligent servant, docile in comparison to the rampant killer robots from blockbuster movies, would be a terrible thing.

Bostrom describes an interesting thought experiment about how to specify the behaviour of a superintelligent machine that produces paperclips. There are at least two undesirable outcomes that stem from "perverse instantiation" of the machine's task. One is that the machine might turn *everything* into paperclips, including humans. Or, faced with a more carefully worded task, the machine first solves the problem of increasing its own intelligence (to become a better paperclip-producer), turning the entire solar system microprocessors. No matter how much we continue specifying our orders, the machine continues to avoid "doing what we mean" in increasingly intricate ways, all with catastrophic outcomes.

I like to think of this idea in terms of Goethe's *Zauberlehrling*,

immortalised by Mickey Mouse in Disney's *Fantasia*. In this tale, the sorcerer's apprentice commands his enchanted broomstick to fill water into a bathtub. Obedient and literal-minded, the magical servant hauls up bucket after bucket, yet continues long after the tub has been filled. Disaster is averted only because the sorcerer himself arrives in the nick of time to save Mickey from drowning. No malice was involved; unlike Bostrom's allegorical owl-chick, the broomstick has no volition other than the dutiful execution of Mickey's task. It was Mickey who failed to be sufficiently precise in how he formulated the task. Had Goethe's broomstick been superintelligent, it seems to me that he might just have killed the apprentice and thrown his body in the tub. After all, 70 percent of it are water, so the task is completed with speed and elegance.

These games of formalism may strike you as trite exercises in deliberately misunderstanding instructions against their intention. However, every programmer or lawmaker knows that this is exactly the reason for why software is so hard to write or laws are hard to formulate. If there is a way to misunderstand your instructions, it will happen.

Thus, there is no need to attribute malice to the superintelligence. Instead, we might be doomed even by a perfectly obedient agent labouring to perform ostensibly innocent tasks. In the words of Eliezer Yudkowsky, "The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else." [1] Of course, a superintelligence with volition, intentionality, or free will, would be even harder to control.

Figure 2: Illustration of *Der Zauberlehrling*. From: *Goethe's Werke*, 1882, drawing by Ferdinand Barth (1842–1892).

[1] Eliezer Yudkowsky, *Artificial Intelligence as a Positive and Negative Factor in Global Risk*, 2006.

The control problem has many aspects, one of which is purely institutional: If the superintelligence is developed in vicious competition between corporations or militaries, none of them is motivated to stop the development, for fear of losing a technological arms race. As a civilisation, we have an unimpressive track record of preventing globally harmful behaviour when individual proximate gains are at hand. This addresses a popular rejoinder to a dystopian intelligence explosion: "Why don't we just pull the plug?" First, it's not clear that there is a plug to pull.[2] Second, who is "we"?

Assuming we can solve the institutional problem of who gets to design the the superintelligence, the solution seems to be to codify our goals and values, so that it won't act in ways that we find undesirable. This, of course, is a problem as old as philosophy: What *are* our values? And again, who is "we": the species, the individual, or our future? Should the paperclip maximiser make sure no children are killed? Should no unborn children aborted? Is it in the alcoholic's interest to receive alcohol or not?

Bostrom speculates if we can "what is good for us" on a higher-order level, asking the superintelligence to extrapolate our desires based on a charitable (rather than literal) interpretation of our own flawed descriptions, possibly in a hierarchy of decisions deferred to ever higher moral and intellectual powers. Assuming that we were

[2] Unlike in the movies, we must assume that a superintelligence can anticipate the efforts of a plucky group of heroes. Bostrom describes a number of ways in which to make yourself immune to such an attack by distributing the computational substrate. Surely a superintelligence can think of more.

to agree on a meaningful "Three Laws of Robotics" that accurately described our values without risking perverse instantiation, how do we then motivate a vastly superiour intellect to follow these laws? I find these speculations both entertaining and thought-provoking.[3]

[3] I particularly like the idea of design an AI with a craving for cryptographic puzzles to which we know the answer.

Some futurists welcome the idea of a paternalistic superintelligence that unburdens our stone-age mind from solving our own ethical questions. Even so, as Bostrom argues, if we share our environment with a superintelligence, our future depends on its decisions, much like the future of gorillas depends on human decisions. If we are currently building the superintelligence, we have but one chance to make it treat us as nice as we treat the gorillas.

The Library of Turing

Bostrom tacitly assumes a strictly functionalist view of the mind. For instance, if we were able to perfectly simulate the neurochemical workings of a brain, this simulation itself would be a mind with consciousness, volition, and agency. Once the functions of the brain are understood, the substrate on which their simulation runs is secondary. From this perspective, questions about cognition, intelligence, and even agency, are ultimately *computational* questions.

Let me invite you to a fictional place that I call the Library of Turing. It is inspired by *La biblioteca de Babel*, described in a short story by Jorge Luis Borges. His library consists of an enormous collection of all possible books of a certain format. Almost every book is gibberish, as had it been typed by intoxicated monkeys. One book consists entirely of the letter A, but another contains *Hamlet*.[4] In Borges' story, the librarians find themselves in a suicidal state of despair; surrounded by an ocean of knowledge, yet unable to navigate it due to the library's overwhelming dimensions. Our brains cannot fathom its size, and our languages cannot describe it[5], "unimaginably vast" does not even come close. However, the metaphor of the library helps us to build some kind of mental image of its size.[6]

[4] A million others contain *Hamlet* with exactly one misprint.

[5] The language of mathematics is an exception. But our brains cannot imagine 10^{25}, the number of stars, much less 10^{123} the number of atoms in the universe. By contrast, the number of books in the Library of Babel is $25^{1,312,000}$.

[6] Daniel Dennett has has used this idea as "the Library of Mendel" to illustrate the vast configuration space of genetic variation. Daniel C. Dennett, *Darwin's Dangerous Idea*, Simon & Schuster 1995.

Behold, then, the Library of Turing. Each volume contains computer code. Much of it is garbage, but some volumes contain meaningful programs in some concrete programming language, let's say Lisp.[7] We can feed these programs to a computer found in each room. Most wouldn't do anything useful, many would cause the computer to crash. From our starting room, lined with bookcases, we see exits to other rooms in all directions. These rooms contain more volumes, more bookcases, and more exits. There is no discernible ordering in the library, but close to where we are standing, a particularly worn volume contains the program

[7] I chose Lisp because of its role as an early AI programming language, but any other language would do. For a cleaner argument, I could have chosen the syntax of Alan Turing's original "universal machine". In the language of computability theory, all these choices are "Turing-complete": they can simulate each other.

```
(print "Hello, world!")
```

which instructs the computer to output the friendly message "Hello world!". Next to it, a book contains the nonsensical phrase "Etaoin shrdlu," not a valid program in Lisp. We shrug and put it back. Thanks to the diligence of some algorithmic librarian, a third book in our room contains the complete code for the primitive natural language processor *Eliza*, written by Joseph Weizenbaum in 1966. This famous little program is able to pretend to have (typed) conversations with you, like this:

"Tell me your troubles."

"My cat hates me."

"What makes you believe cat hates you?"

The primitive grammatical mistake of leaving out the pronoun gives you a clue that *Eliza* works by very simple syntactic substitutions, without any attempt at understanding what you mean. Still, it's better than just saying "Hello, world!" If we continue browsing, we might find the code of some natural language processors that are state-of-the-art in the 2010s, such as IBM's *Watson*, or Apple's *Siri*.

Indeed, our conceit is that every algorithm appears somewhere in Turing's library. [8] In particular, if we accept functionalism, some book describes the behaviour of Einstein's brain: On input "What is $5 + 5$" or "Do you prefer Brahms?" or "Prove the Riemann hypothesis" it will give the same answer as Einstein. We have no idea how this program looks. Maybe it is based symbolic artificial intelligence, like *Eliza*. Or maybe the bulk of the books is taken up by a "connectome" of Einstein's actual brain, *i.e.*, a comprehensive map of its neural connections. Such a map has been known for *C. elegans*, a transparent roundworm with 302 neurones, since the mid-1980s, and even though Einstein's brain is considerably larger, this is not absurd in principle.

```
(defineq

(doctor
 (lambda nil
  (prog (sentence keystack
phraselist)
     (setsepr "" " " " ")
     (setbrk "." "," ? | - + "("
")" L32 @ BS L14)
     (setq flipflop 0)
     (control t)
     (sentprint (quote (tell me your
troubles"."
please terminate input with an
enter)))
     (setnone)
   a     (prin1 xarr)
     (makesentence)
      .e
     bag)
    (nil)
    (do you say 1 2 3 4 for some
special reason)
    (what might 1 have to do with
your problem)
    (do you often say ' 2 3 4 ')
    (perhaps you feel that you bite 3
4))
   zzyyxx)
  rules)

 (defprop oh 10 priority)

 (defprop oh (((0 my oh my 0)
    (nil)
    (pre (1 my-oh-my 5)
     zzyyxx))
   ((0 oh my 0)
    (nil)
    (pre (1 oh-my 4)
     zzyyxx))
   zzyyxx)
  rules)

 (defprop zzyyxx (((0)
    (nil)
    newkey))
  rules)

stop
```

Figure 3: Fragments of the source code of *Eliza*, an early natural language processor. The whole program takes just over 1000 lines.

[8] Some programs may be so big as to be spread over several books, maybe referring to other volumes as subroutines much like modern software is arranged into so-called software libraries.

SOMEWHERE IN TURING'S LIBRARY there must be description of a computational process that is vastly superior to every human brain. Otherwise Einstein's brain would have been the smartest algorithmic intelligence allowed by the laws of logic, or close to it. Yet it seems unlikely, not to say self-congratulatory, that evolution on the planet Earth succeeded in constructing the ultimate problem solving device in just a few million generations.

Translated into this setting, Bostrom suggests that our discovery of this description would have catastrophic consequences. The library contains a monster. Our exploration should proceed with prudence rather than ardour.

Searching for the Monster

Given that the description of a superintelligence exists somewhere in the library, all that remains is to find it. If you don't work in computer science, this task seems to be relatively trivial. After all, even a mindless exhaustive search will find the volume in question. While this approach is not a particularly tempting exercise for a human, computers are supposed to be good at performing mindless,

well-defined, repetitive, and routine tasks at high speed and without complaining. Thus, sooner or later we stumble upon the monster; the only debate is whether the search time is measured in years, generations, or geological time scales.

But this is a fallacy. It is a fallacy because of the unreasonable growth rate of the exponential function, and the powerlessness of exhaustive search.

To appreciate this, consider again the 23 symbols that make up our friendly "Hello World!" program. We could have found it, *in principle*, by exhaustively searching through all sequences of characters that make up Lisp programs.

How much time does this take? That is a simple exercise in combinatorics, after we fix some details—how many symbols are there, how fast is the computer, etc. But no matter how you fix these details, the result is disappointing.[9] If a modern computer computer had started this calculation when the universe began, it would have gotten to somewhere around `(print "Hello,")`. You could gain a few extra letters by throwing millions of computers at the problem. But even then, the universe does not have enough resources to allow for an exhaustive search for even a very simple program. Either you run out of time before the universe ends, or you run out of protons to build computers with. Computation is a resource, exponential growth is immense, and the universe is finite.

[9] Lisp is written in an alphabet of at most 128 characters. A modern computer with 10^9 operations per second, running since the Big Bang, would not have finished all 128^{13} strings of length 13.

Bostrom pays little attention to this problem. For instance, in his discussion of current efforts on brain simulation, he writes:

> Success at emulating a tiny brain, such as that of *C. elegans*, would give us a better view of what it would take to emulate larger brains. At some point in the technology development process, once techniques are available for automatically emulating small quantities of brain tissue, the problem reduces to one of scaling.

Well, to me, scaling *is* the problem.

One might object that this observation is a mundane, quantitative argument that does not invalidate the compelling fact that in principle, the exhaustive search will sooner or later stumble upon the fully-fletched "Hello World!"-program, and eventually the monstrous superintelligence. But that is exactly my point: You can simultaneously accept machine-state functionalism *and* be blasé about the prospects of AI. Speculation about the imminent discovery of the monster in Turing's Library must be grounded in computational thinking, which is all about the growth rates of computational resources.

Could there be another way of discovering the superintelligence than exhaustive search? Certainly. After all, nature has discovered one of the monsters, the brain of *Homo sapiens*, starting with very simple "brains" hundreds of millions of years ago, and refining them stepwise by natural selection and mutation in environments that favoured cognition. Thus, there is some gradient, some sensible set of stepwise refinements, that Nature has followed through the Library

of Turing to arrive at Einstein's brain. [10] We just don't know how to recognise, nor efficiently follow this gradient. In the enthusiasm surrounding artificial intelligence in the 1960s we might have thought that we were on such a gradient. If so, we have abandoned it. The current trend in artificial intelligence, away from symbolic reasoning and towards statistical methods like machine learning, that do not aim to build cognitive models, strikes me as an unlikely pathway.

[10] If we entertain this idea, then a pathway to intelligence would be the simulation of competetive environments in which to artificially evolve intelligence by mutation and selection. This replaces the problem from simulation of entire brains to simulation of entire adaptive environments, including the brains in them.

Given our current understanding of the limits of computation, there are many relatively innocent-looking computational problems that are computationally hard.[11] After more than a generation of very serious research into these problems, nobody knows anything better than exhaustive search for many of them. In fact, while the amazing features of your smartphone may tell you a different story, the main conclusion of a generation of research in algorithms is bleak: for most computational problems we don't know what to do.[12]

[11] This claim is formalised in various hypotheses in the theory of computational complexity, such as "P is not NP" and the Exponential Time Hypothesis.

[12] Not even functioning quantum computers would change this.

Let me repeat that I do not try to rule out the *existence* of a monster. I merely point out that even though a monster exists, we may never run into it. Turing's Library is just too vast, nobody has labelled the shelves, much less put up useful signs and maps. Our intuition is mistaken when it extrapolates that we will soon have mapped the entire library, just because we've seen more of it than our grandparents ever did. Moreover, imminent and rampant progress in exhaustive search may have been a reasonable hope half a century ago. But today we know how hard it is.

Research Directions

If the algorithmic perspective above is correct, then there is nothing to worry about. The superintelligence is an entertaining fiction, no more worthy of our attention than an imminent invasion by aliens,[13] eldritch gods from the oceans, or a sudden emergence of magic. The issues raised are inherently unscientific, and we might as well worry about imminent overpopulation on Venus or how many angels can dance on the head of a pin.

[13] Here is the analogy in full: We have explored the Moon and are about to explore Mars. We should worry about imminent contact with alien civilisations.

But I could be wrong. After all, I wager the future of human civilization on a presumption about the computational complexity of an ill-defined problem. I do think that I have good reasons for that, and that these reasons are more firmly grounded in our current understanding of computation than Bostrom's extrapolations. However, if I'm wrong and Bostrom is right, then humanity will soon be dead, unless we join his research agenda. Thus, from the perspective of Pascal's wager, it may seem prudent to entertain Bostrom's concerns.

The problem domain opened up by *Superintelligence* may lead to valid research in various areas, independently of the plausibility of the underlying hypothesis.

For instance, the problem of codifying human values, a formalization of ethics, is a problem as old as philosophy. Bostrom's prediction injects these questions with operational significance, as these codified ethics would be the framework for the incentive structure needed to

solve the control problem. However, an "algorithmic code of ethics" is a worthwhile topic anyway. After all, our digital societies are becoming increasingly controlled by algorithmic processes. These processes are nowhere near superintelligent, but they are highly influential: Algorithms choose our news, decide our insurance premiums, and filter our friends and potential mates. Their consistency with human values is a valid question for ethicists and algorithmicists alike, no matter if the algorithms are authored by Google's very real software engineers or a very hypothetical paternalistic AI.

Another issue related to the control problem is the ongoing work on the semantics and verification of programming languages. This has been a well-established subfield of computer science for a generation already. Some of the these issues go back to Gödel's incompleteness theorem's from the 1930s, proving that formal systems are unable to reason about the behaviour of formal systems. In particular, it has turned out to be a very difficult problem of logic to specify and verify the intended behaviour of even very short pieces of computer code. It may well be that questions about harnessing the behaviour of a superintelligence can be investigated using these tools. Other scientific areas may provide similar connection points: Incentive structures are a valid question for game theory and algorithmic mechanism design, which resides on the intersection of computer science and economics. Or the computational aspects of artificial intelligence may become sufficiently well defined to become a valid question for researchers in the field of computational complexity. This might all lead to fruitful research, even if the superintelligence remains eternally absent. But whether superintelligence becomes a valid topic for any of these disciplines is very unclear to me. It may depend on purely social issues within those research communities.

It may also depend on the availability of external funding.

This last point leads me to speculate about a very real potential impact of Bostrom's book. It has to do with the perception of artificial intelligence in the public eye. Research ethics are a political question ultimately decided by the public. The most Herculean promise of current artificial intelligence research is the construction of general artificial intelligence. What if the public no longer saw that as a goal, but as a threat? If Bostrom's dystopian perspective leads to increased public awareness about catastrophic effects of superintelligence, then general AI would not appear under "Goals" in a research plan, but under "Ethical considerations" that merit to public scrutiny along with climate change, sensitive medical data, rampant superviruses, animal testing, or nuclear disasters. And these, largely political, requirements may themselves necessitate research into the control problem.

We may want to ponder the consequences of running into a monster in the Library of Turing no matter how plausible that event may be.