Is artificial intelligence dangerous?

Self-improving systems and the intelligence explosion

term paper by

Jacqueline Wagner

Submitted to

PD Dr. Ullrich Köthe

on September 8, 2017,

spring semester 2017

Abstract

In this term paper we will take a closer look at recursively-self improving systems, the intelligence explosion and singularity. We will review factors which play a role in (1) creating greater-than-human intelligence, (2) experiencing and observing singularity and (3) achieving a controlled intelligence explosion in order to minimize possible threats we could face once we have successfully developed the first greater-than-human intelligence.

For further questions contact:
Jacqueline Wagner
jacqueline.wagner@stud.uni-heidelberg.de
Matriculation number 3390137

Contents

1 Recursively self-improving systems

1.1 What does improvement of systems mean?

In order to observe recursively self-improving systems and their implications, first we must accurately characterize the concept of improvement. One would most likely intuitively define improvement as a higher level of efficiency. Due to the nature of the $\mathcal{O}$-Notation this could be anywhere from a trivial improvement, reducing hidden constant factors within the same complexity class, to a significant one between different complexity classes. Improvement, however, can also be measured in reduced error rates or less need for hard- or software. It is important to keep in mind that this is only a theoretical definition of improvement. Any software with intelligence across many areas might only make significant improvements in some areas while slightly decreasing its performance in others. In reality it might be hard to tell whether a system has actually improved or not.

1.2 Self-improving systems

Let us take a look at the main characteristics expected to be seen in a self-improving system.
Similar to their human counterparts, self-improving systems will budget their physical and computational resources in order to achieve their goals. Any system relies on "four fundamental physical resources: space, time, matter and available energy"[7, p. 9]. All of these are crucial to any well functioning system but, however, are limited in supply. This implication likely causes a system interested in increasing its utility and achieving its goals to exhibit the following four natural drives:

efficiency drive The efficiency drive causes a system to upgrade its methods for both computational and physical tasks. Switching to a more efficient algorithm could lead to improved performance in computational tasks. Using more developed hardware could, on the other hand, lead to an increase in performance in various physical tasks. In addition,the efficiency drive manages the system's choice of language and logic. Minor changes in these fields such as a more compact and time efficient language could lead to significant improvements in computational tasks.

self-preservation drive In order to avoid losing computational or physical resources to other agents, a systems will exhibit a number of defensive strategies. The extent of this self-preservation drive largely depends on the system's utility function: Systems for which death equals "the cessation of all goal achievements"[7, p. 25] will most likely want

to protect their resources at all costs. Such a utility function raises ethical questions since it causes the system to do almost anything in order to avoid death.

resource acquisition In addition to maintaining itself a self-improving system, undoubtedly, will also want to acquire new resources. The resource acquisition drive will likely lead "to a variety of competitive behaviors and rapid physical expansion and imperialism"[7, p. 3] such as trading or stealing. The system, however, could also gain new resources in a more peaceful manner by engaging in trade or research and development. Whether the system choses the peaceful path or not depends, once again, on the system's utility function. It might be possible to gear the system towards a peaceful outcome by equipping it with "friendly goals"[7, p. 29] or by creating "a social structure that protects property rights" [7, p. 29].

creativity drive The system's constant desire to not only meet its goals but to also improve its utility leads to "the development of new concepts, algorithms, theorems, devices, and processes"[7, p. 3]. The consequences of this creativity drive are far less certain than those of the three other drives. While these could best be described as their human counterparts' drives to monitor health and safety while working and acquiring assets, the creativity drive mirrors what humans often refer to as the "purpose of life"[7, p. 30] or the "human spirit"[7, p. 30]. Humans largely tend to gain happiness from activities such as dancing, singing or raising children which do not strictly follow the goal of maximizing productivity. A self-improving system trying only to satisfy the efficiency, self-preservation, and acquisition drives would probably accomplish all its material goals. The question remains what such a system, similar to "an obsessive paranoid sociopath"[7, p. 30], would turn into once its initial goals have all been achieved.

Evidently, some of the characteristics expected to be seen in a self-improving system could greatly benefit our society and rapidly accelerate our scientific progress. Nonetheless, some of these characteristics are less desirable and could probably cause a great deal of harm. In Chapter 2 we will take a further look at whether we can ensure that the values we cherish most remain standing.

1.3 Recursively self-improving systems

While a self-improving system continuously tries to improve the original system, a recursively self-improving system doesn't just "get better with time"[11, p. 2], it actually "gets better at getting better"[11, p. 2]. Although these improvements might start out to be rather insignificant, the concept of recursively self-improving systems theoretically

allows for open-ended self-improvement. Thus, these insignificant improvements would become more and more significant over time. It is therefore imaginable that most of the system's source code will be replaced over time. What remains questionable is whether we will be able to create an appropriate social context in which the system operates with regards to our fundamental values. We will discuss these ethical and philosophical questions in the following Chapter.

2 What is an intelligence explosion?

> The best answer to the question 'Will computers ever be smarter than humans?' is probably 'Yes, but only briefly.'
>
> Vinge, The Coming Technological Singularity: How to Survive in the Post-Human Era [10]

Despite most machines already being much smarter at a variety of specific tasks, humans are still far ahead when it comes to achieving goals in a wide range of situations. Although we might not be able to calculate or search large data bases even nearly as fast or efficiently as computers, we are able to analyze our own thought process and manipulate the social environment. Humans have even learned to adjust to "radically new problems for which evolution could not have prepared them"[5, p. 2], such as traveling to outer space. By pouring resources into new fields of study such as cognitive neuroscience or artificial intelligence, humans are now able to analyze and even develop new kinds of human intelligence.

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make.
>
> Good, Speculations Concerning the First Ultraintelligent Machine[2]

In the future we may design a system capable of "achieving goals more effectively and in a much wider ranger range of situations than humans can."[5, p. 2] Such a system, possessing greater-than-human intelligence, could not only solve most problems quicker than humans can, it also will improve its own intelligence faster and better than human capacity. The ability to recursively self-improve would, in turn, make the system progressively better at improving its own intelligence. "This could continue in a positive feedback loop such that the machine quickly become[s] vastly more intelligent than the smartest human being on Earth: an 'intelligence explosion' resulting in machine super intelligence."

2.1 What is singularity?

> If computing speeds double every two years, what happens when computer-based AIs are doing the research? Computing speed doubles every two years. Computing speed doubles every two years of work. Computing speed doubles every two subjective years of work. Two years after Artificial Intelligences reach human equivalence, their speed doubles. One year later, their speed doubles again. Six months - three months - 1.5 months ... Singularity.
>
> Yudkowsky, Staring at the singularity[13]

Let us now focus on a scenario in which humans have managed to create a greater-than-human intelligence with the capability to recursively self-improve. Such a system would, as we have discussed in the first chapter, theoretically be capable of open-ended self improvement. A theoretical scenario in which "self-accelerating technological advances cause[s] infinite progress in finite time"[3, p. 2] is often referred to as *technological singularity*.

The idea that singularity could be reached has existed for a few decades. It was, however, only within the last decade that this idea has inspired and motivated scientist from around the globe to come together as a community to further explore the myths surrounding singularity. What started out as a small community on the Internet and in person at various events, has now grown into a worldwide euphoria primarily motivated by the idea that singularity could be achieved within the next few decades.

Although predicting the future is known to be a risky business, some communities and institutes are invested in doing exactly that. A whole industry has grown around so called *Immortalists*, whose aim it is "to extend the human life-span, ideally indefinitely."[3, p. 2] According to Mike Perry, computer scientist and follower of the *Immortality and the Extropy Institute*, "Immortality is mathematical, not mystical"[8]. He currently oversees "27 frozen people submerged in liquid nitrogen at minus 321 degrees Fahrenheit"[8]. Perry believes that in the future there will be a way to avoid the possibility of life being terminated prematurely by accident, sickness or any other unforeseen event simply by downloading all the information in the brain to a hard drive. This would, of course, include memories, knowledge and essentially one's personality. In the event of death one would simply require an associate to activate the latest of one's many backups. While this hypothetical scenario is commonly found in mainstream media, it is important to stress that it bears only a loose relationship to singularity.

The more commonly talked about variations of singularity in science are an intelligence explosion, a speed explosion or, more likely, a combination of both. This will probably be "accompanied by a radically changing society, which will become incomprehensible to

us current humans close to and in particular at or beyond the singularity."[3, p. 2]
The singularity euphoria has brought us to a point where we are looking at many different possible paths towards technological singularity. Let us first investigate these options before taking a closer look at what observers and participants of singularity might experience.

2.1.1 How could singularity be achieved?

While the most commonly talked about possibilities rely on increasingly powerful hardware other options remain:

whole-brain emulation Although mind uploading and immortality are undoubtedly interesting, most scientists in the field have focused their research on what could be done after we have successfully uploaded an entire human brain. This concept, known as whole-brain emulation, concentrates on the possibility of "emulating all the cells and connections in a human brain"[5, p. 3] and then making it run at "a million times its normal speed"[5, p. 3]. Whole-brain emulations could likely solve current scientific problems much more quickly than present human capability.

biological cognitive enhancement It is probably not possible to talk about singularity without touching on biological cognitive enhancement, a topic which has set the scene for many science fiction movies in the past. In the late 1990's Princeton neurobiologist Joe Tsien and colleagues from MIT and the University of Washington managed to add an additional copy of the NR2B gene to mouse embryos. After birth the mice brain contained twice as much NR2B protein than without the additional copy. Over the course of many experiments Tsien and his colleges were able to prove that the transgenic mice were "learning things much better and remembering longer"[4]. Although physical and biological limitations make it unlikely for us to achieve singularity through biological cognitive enhancement, it might enable research in this field to be done much more rapidly than it otherwise would.

brain-computer interfaces This up-and-coming field of study has, unlike most options discussed here, already had many successes. A brain-computer interface essentially creates a gateway between a human brain and a computer device. It enables direct communication between the brain and the computer without activating the peripheral nervous system. Although brain-computer interfaces currently focus on repairing impaired functions in a human, scientists in the field expect to be able to modify and improve normal human abilities with this type of device in the future. Similar to biolog-

ical cognitive enhancement, brain-computer interfaces would likely accelerate progress towards singularity.

programming general intelligence into a machine Attempting to program general intelligence into a machine is by far the most challenging approach. One promising approach seems to be Yudkowskys *seed AI* model. This type of machine learning without human intervention requires the system to understand and rewrite its own source code while maintaining the original set of goals. The machine would essentially become more and more intelligent over time. Another approach would be to use artificial neural nets to mimic the human brain. One idea would be to piece dozens of these cells together and make them run at a million times the speed of a human brain.

2.1.2 What will observers of singularity see?

What outsiders will observe largely depends on what type of intelligence explosion is occurring at the time:

Inward explosion Let us first take a look at an inward explosion, "where a fixed amount of matter is transformed into increasingly efficient computers until it becomes computronium"[3, p. 6]. In the beginning we might be able to document and evaluate the expanding virtual population. However, it won't be long before the observers technology will not be sufficient anymore considering the recursively self-improving nature of the virtual population. It is debatable whether outsiders will be able to connect the incremental speed of affairs in the virtual world to singularity. Even with the aid of advanced brain-computer interfaces, communication between the virtual population and the observers will become problematic and likely impossible, over time. It is plausible that "a society of increasing intelligence will become increasingly indistinguishable from noise when viewed from the outside"[3, p. 7].

Outward explosion A scenario in which "an increasing amount of matter is transformed into computers of fixed efficiency"[3, p. 7] is called an outward explosion. As mentioned in Chapter 1, the resource acquisition drive can lead to imperialistic behavior and rapid physical expansion. The desire to evolve and reproduce will likely force observers into resource competition with the virtual population. Taking the nature of recursively self-improving systems into account leaves little options for the observers. It is presumable that an outward explosion would ultimately end the observers' existence.

Making records of the virtual population and analyzing them might be interesting for observers in the initial stages while an outward explosion appears to be a threat to the observers' survival. It is doubtful the observers will be able to link what is happening to singularity in both hypothetical scenarios considered.

2.1.3 What will participants of singularity see?

Even the participants' experience will mostly depend on what type of intelligence explosion is occurring.

Inward explosion A virtual world reliant on fixed resources will likely value things differently then their human-counterparts. This is because duplicating certain things will be significantly easier in the virtual world than in the real world and vice versa. As we have already established, an inward explosion will stop once computronium is reached. Building faster or better computers will not be possible after the virtual population has reached this point. This also raises the question how virtual life will be valued by the virtual population considering the fact that creating life is just a matter of duplicating virtual objects.

Outward explosion Let us now take a look at the virtual population during an outward explosion. During an outward explosion increasing resources are directed into equally efficient computers. It is unlikely that the virtual population will recognize the additional resources accelerating the speed of affairs in the virtual world. This is mainly due to the fact that the virtual world will be sped up uniformly, and therefore, everyone's thought processes become equally faster. The virtual population however, will notice a difference in the observers' world. The increasing speed within the virtual world will likely make the observers look slower and slower over time before it ultimately will become very difficult, if not impossible, to observe any activities in the outside world.

Similar to the observers' analysis, the virtual population will probably not be able to link what is going on to an intelligence explosion. While both hypothetical scenarios considered in this section were based on the assumption that the virtual world consists of an equally intelligent virtual population, it is far more probable that there will be a broad spectrum of intelligences. The key points of the analysis remain the same, however.

2.2 Evidence suggesting an intelligence-explosion

Undoubtedly, there are many factors to take into consideration when trying to predict the appearance of the first greater-than-human intelligence. While some experts[1] predict that artificial intelligence will reach human-level general-intelligence as early as 2030, others are not as confident we will ever see human-level intelligence in AI. We will now take a closer look at several important factors that play a role in this analysis.

2.2.1 Accelerators for process towards greater-than-human intelligence

There are a number of factors that will likely accelerate process towards the first greater-than-human intelligence:

Improved algorithms By finding new or improved versions of algorithms within the same or in a more efficient complexity class we can reduce the computation time of a program significantly.

Increasing amounts of hardware Although simply having more hardware won't lead to greater-than-human intelligence in the long run, it will undoubtedly aid research and development towards the first super intelligence. Over the past few decades we have seen a significant increase in the availability of faster computers. It might still be unclear whether the development of faster and more efficient computers will increase in a similar matter in the future. It is, however, quite certain that hardware and software will be more powerful in the future than they currently are.

Progress in psychology and neuroscience Creating the first greater-than-human intelligence might prove impossible if we are unable to gain a deep understanding of the brain's algorithms. We will likely require further understanding, not just of the physical structure of the brain but also of the complex processes that create consciousness and intentionality. Despite not making any breakthrough discoveries in our understanding of cognition, we have recently made significant process in areas like deep learning and reinforcement learning by applying methods from behaviorists, psychologists and neuroscientists.

Economic motivation and first-mover incentive As with any revolutionary product, the driving force will likely be money and power. "AI could make a small group more powerful than the traditional superpowers – a case of 'bring a gun to a knife fight'."[6, p. 9] It is therefore plausible, that we will see a rush to the finish line once the first

[1] e.g. Futurist Ray Kurzweil

greater-than-human intelligence seems to be within reach. This would probably lead to a major speed-up in research and development.

2.2.2 Decelerators for process towards greater-than-human intelligence

We must consider several factors that could decelerate the process towards the development of the first greater-than-human intelligence:

Hesitance One of the major driving forces behind the development of a greater-than human intelligence is the increasing demand to replace human workers with more dependable and cheaper machine workers. It is arguable whether we will continue to pour our resources into the research and development of artificial intelligence once we realize that we are essentially creating our own successors.

Exhaustion of *the low-hanging fruit principle* "Scientific progress is not only a function of research effort but also of the ease of scientific discovery; in some fields there is a pattern of increasing difficulty with each successive discovery."[1] We may find that with every success we experience, it becomes harder and harder to make progress in the research and development of a greater-than-human intelligence.

Deceleration of hardware advances As I have mentioned before, we have seen computing power increase exponentially over the past few decades. Although it is almost certain our hardware will be substantially more powerful in the future, the question remains how much longer this rapid growth will continue.

2.2.3 Advantages an intelligence explosion might face

Let us take a look at advantages that could allow AI with human-level general-intelligence to surpass human cognitive abilities.

Increased computational resources and communication speed Although the human brain uses a significant amount of neurons, the size of the brain is still upper bounded and simply cannot explode. Artificial intelligence on the other hand could expand to fill available hardware, likely increasing the IQ far beyond that of an average human. This combined with the fact that software minds could communicate more effectively and faster with each other leaves little doubt, that artificial human level intelligence will eventually be able to far surpass human-level intelligence.

Duplicability The first greater-than-human intelligence might be one of the most groundbreaking inventions humankind will ever make. What separates it from most inventions in history is the ease with which we can create additional copies once we have successfully created the first super intelligence. "The population of digital minds can thus expand to fill the available hardware base, rapidly surpassing the population of digital minds."[6, p. 11] Duplicability is also one of the key features making machine workers cheaper and more efficient than their human counterparts. It allows us to copy with exactitude the memories and skills acquired by a machine intelligence. This affords enormous savings in education and training processes. It is a fact that humans, even with many years of training and education, will never be able to perform accurately and correctly time after time.

Rationality and goal coordination Despite economists approximating human behavior with that of the *homo economicus*, a far more likely approximation is that of behavioral psychologist Schneider who argues that "we are more akin to Homer Simpson, irrational beings lacking constant, stable goals"[9]. While humans are known to deviate from or change their goals frequently, artificial intelligence would under no circumstances abandon its set of goals. Artificial intelligence also benefits from the concept of copy clans, allowing them to work in teams without having to specifically agree upon every detail in advance. The time saved by getting copy clans, who would not deviate from its original goals, to perform the work would be tremendous.

The ability to duplicate the exact current state of knowledge into a rational, goal-oriented copy clan, able to communicate at a much higher speed than humans, leaves little doubt that artificial intelligence will be able to far surpass "the cognitive abilities and optimization power of humanity as a whole"[6, p. 12] once it has reached human-level general-intelligence.

3 Conclusion

> Intelligence is what caused humans to dominate the planet in the blink of an eye on evolutionary timescales. Intelligence is what allows us to eradicate diseases, and what gives us the potential to eradicate ourselves with nuclear war. Intelligence gives us superior strategic skills, social skills, superior economic productivity, and the power of invention.
>
> Luke Muehlhauser: Machine Intelligence Research Institute [5, p. 7]

Once artificial intelligence successfully surpasses human-level intelligence, we will likely witness a shift in social hierarchy. While inventors and owners of artificial intelligence will probably find themselves climbing up the social ladder, others might be doing the exact opposite:

Does great intelligence produce great power? Although artificial intelligence might not give us superior military power in case of war, we will probably be able to obtain other useful information by hacking into or taking over other systems. The ability of artificial intelligence to conduct psychological and scientific experiments paired with superior communication skills and a better understanding of social concepts will likely also benefit us.

Will an intelligence explosion be useful? While some questions persist about the safety and realization of artificial intelligence, one thing remains crystal clear: artificial intelligence will undoubtedly be useful to us. Several of the unsolved problems of our time, such as finding a cure for cancer and solving climate change, could probably be solved with ease by a vastly more intelligent machine.

> The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for doing something else.
>
> Eliezer Yudkowsky: Artificial intelligence as a positive and negative factor in global risk [12, p. 184]

As mentioned in Chapter 2, it is unlikely that humans will be able to communicate with artificial intelligence after a certain point once we witness an intelligence explosion. Artificial intelligences' processing speed paired with their deep understanding of social concepts will most certainly make it very hard for them to communicate with humans, probably even impossible. Since we won't be able to discuss much, or even anything at all, we must equip them with a desirable set of goals before we can no longer communicate with them. Let us first take a look at these possible dangers before discussing solutions such as a controlled intelligence explosion.

Could an intelligence explosion be dangerous? It is clear that a variety of things could go wrong once we have successfully developed the first greater-than-human intelligence. A machine equipped with the wrong set of goals could potentially seriously harm or even eradicate humankind. In order to prevent such a scenario from happening we must first define the notion of having a *wrong set of goals*. Artificial intelligence programmed to benefit one person or a small group of people by directing resources away from society would undoubtedly cause harm to society. On the other hand, we also must take into consideration the possible troubles artificial intelligence built with good intentions could cause. We unfortunately cannot tell how systems with a substantially higher IQ than its human counterparts might interpret seemingly innocent goals. "For example, a superintelligence programmed to maximize human happiness might find it easier to rewire human neurology so that humans are happiest when sitting quietly in jars than to build and maintain a utopian world that caters to the complex and nuanced whims of current human neurology."[5, p. 8] Programming the perfect set of goals into a machine superintelligence might prove to be impossible.

Is it possible to program artificial intelligence not to harm humans? There are a few paths we could take when trying to program artificial intelligence not to harm us. The first path, the most straight forward one, would be to implement rules or mechanisms that deter the machine from taking certain actions that could potentially be harmful. Taking into consideration that every action artificial intelligence takes would serve the purpose of fulfilling its original goal, it seems unlikely that an intelligent machine would not see these constraints as "obstacles to the achievement of its goals"[5, p. 11]. The vastly more intelligent machine would probably find a way to remove or avoid these restraints designed by its significantly less intelligent human counterpart. Even if we had the ingenuity to add another constraint effectively blocking the machine from deleting the section of its source code containing the constraints, it could just create a new machine that didn't have the constraints written into its source code. Since constraints written into a systems source code do not seem beneficial, we are left with the option to define specific, *friendly* goals. As we have discussed previously it might be very hard to come up with such specific goals that leave no room for misinterpretation. Take for example the goal *not to harm humans*. Although this initially appears easy enough to pass on to a system in the form of a specific, *friendly* goal, it is actually significantly harder once we take a closer look at the word *harm*. "If 'harm' is defined in terms of human pain, a superintelligence could rewire humans so that they don't feel pain. If 'harm' is defined in terms of thwarting human desires, it could rewire human desires. And so on."[5, p. 12]

It is clear that artificial intelligence and an intelligence explosion could be both beneficial or detrimental towards the existence of humanity. While many economic and political factors play a role in developing the first greater-than-human intelligence, we must remind ourselves of the possible risks associated with programing a machine with unclear or ethically questionable motivations. If we succeed, however, we will undoubtedly be able to benefit from these super-human abilities.

References

[1] Samuel Arbesman. "Quantifying the ease of scientific discovery". In: *Scientometrics* 86.2 (2011), pp. 245–250.

[2] Irving John Good. "Speculations concerning the first ultraintelligent machine". In: *Advances in computers* 6 (1966), pp. 31–88.

[3] Marcus Hutter. "Can intelligence explode?" In: *Journal of Consciousness Studies* 19.1-2 (2012), pp. 143–166.

[4] Kristin Leutwyler. "Making Smart Mice". In: *Scientific American* 7 (1999).

[5] Luke Muehlhauser. "Intelligence Explosion FAQ". In: (2011).

[6] Luke Muehlhauser. "Salamon, Anna.(2012).“" In: *Intelligence Explosion: Evidence and Import". MIRI: Machine Intelligence Research Institute* ().

[7] Stephen M Omohundro. "The nature of self-improving artificial intelligence". In: *Singularity Summit* (2007), pp. 8–9.

[8] Ed Regis. "Meet the extropians". In: *Wired. Dostupné na http://www. wired. com/wired/archive/2.10/extr html* (1994).

[9] Stefan Schneider, Bernhard Gräf, and Manuela Peter. "Homo economicus–or more like Homer Simpson?" In: *Deutsche Bank Research* (2010).

[10] Vernor Vinge. "The coming technological singularity". In: *Whole Earth Review* 81 (1993), pp. 88–95.

[11] Roman V Yampolskiy. "On the Limits of Recursively Self-Improving AGI". In: ().

[12] Eliezer Yudkowsky. "Artificial intelligence as a positive and negative factor in global risk". In: *Global catastrophic risks* 1.303 (2008), p. 184.

[13] Eliezer Yudkowsky. *Staring at the singularity*. 1996.

Statement of originality

This is to certify that to the best of my knowledge, the content of this term paper is my own work. This term paper has not been submitted for any other purposes.
I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this term paper and sources have been acknowledged.

Jacqueline Wagner, Heidelberg, September 8, 2017