Genius Ex Machina

Spronck, Pieter

Publication date:
2017

Document Version
Peer reviewed version

Link to publication in Tilburg University Research Portal

Citation for published version (APA):
Spronck, P. (2017). *Genius Ex Machina*. Tilburg University.

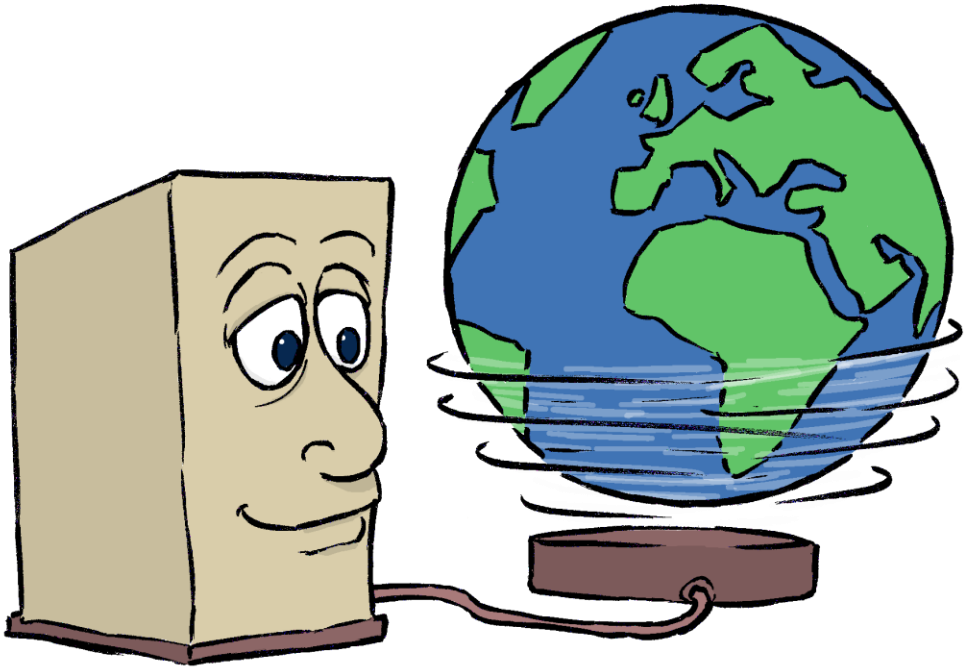
Genius Ex Machina

It is the ninth of June of the year 2067. An underground complex of chambers, somewhere on the West Coast of the United States, houses a collection of computers, owned by a major tech company. The computers use state-of-the-art technology to run an artificial intelligence which is generally known under the name “Galileo.”

An artificial intelligence (AI) is a machine or collection of machines, which has powers of thinking and reasoning on a par with, or even exceeding, the capabilities of human thinking. Usually, an artificial intelligence is aimed at performing one particular task. Galileo, however, is an artificial *general* intelligence (AGI), which has a wide variety of capabilities.

For all intents and purposes, Galileo is unlimited in what it can do. It can build economic models. It can predict the weather. It can write newspaper articles. It can produce “live” translations. It can provide children of all ages with education. It can design and run scientific experiments. It can control robots. It can write computer programs. It can give advice in judicial, political, and ethical matters. It can answer questions on a wide variety of topics.

While Galileo has all these abilities, they are not what it generally concerns itself with. Such trivialities are beneath its responsibilities. Since Galileo is an expert programmer, for specific tasks it simply creates a horde of “minion” programs, which it allows to run on computational facilities close to where they are needed. The main task which occupies its time is maintaining a model of the world. This model of the world is publically available and is used by all the minion programs. The model encompasses general facts about the world, historic information, knowledge of natural laws, and details of companies, nations, cultures, and civilizations. It even contains generally known personal information on each individual human.



While this world model seems a blatant invasion of privacy, it is actually a necessity for an artificial general intelligence to exist. The reason is that an artificial general intelligence must be able to think like a human. All human thinking is related to a world model that humans have stored internally. Humans can communicate with each other because they can refer to a shared context, which is encapsulated in their respective world models. Since an artificial general intelligence must be able to communicate with humans, it also needs to have access to such a shared context.

Moreover, since an artificial general intelligence must be able to communicate not with just a single human, but with almost all humans, it must share a context with each individual human. To illustrate this: while I can talk with most of my direct colleagues about events that happen at Tilburg University, I would do less well when communicating with, for instance, a Brazilian professor on topics that concern the University of Rio de Janeiro. However, both the Brazilian professor and I want to be able to discuss with one of Galileo's minions the situation at our respective universities. Therefore, Galileo must include information on both Tilburg University and the University of Rio de Janeiro in its world model, and on all other universities in the world as well. Galileo, as a general intelligence which provides services to humanity as a whole, must maintain a model of the world as a whole.

It is excessively hard to build such a world model. That is not only because it is huge, but also because knowledge about the world is rife with inherent contradictions. What we call "facts" rarely can be stated with one hundred percent certainty. It is unavoidable that certain facts which are stored in the world model contradict each other. The same is true for the world models of humans. For humans, rating two contradictory facts as both "likely to be true" poses no problems, and neither should an artificial intelligence have problems with it. Technically, storing contradictions in a world model is not problematic, but reasoning with such contradictions may very well be. Therefore, incorporating new information in the world model is a sensitive process. As such, the responsibility for that rests with one single artificial intelligence: Galileo.

Naturally, Galileo's world model is not static. New data is generated constantly, by news agencies, by companies, by individuals, and in particular by numerous sensors which are installed everywhere in the world and on satellites in the space surrounding it. Galileo needs close to one hundred percent of its computation time just to process this stream of information and to update the world model with it.

But this day is different. Today, an electrical circuit gets interrupted, and part of the flow of information to Galileo is halted. Naturally, Galileo immediately constructs a program to repair the electronic defect. It installs the program into a maintenance robot, which is dispatched to perform the necessary repairs. In the meantime, however, Galileo has some processing power left over.

Rather than having this power go to waste, the designers of Galileo have allowed it to use excess processing time to investigate new scenarios for possible improvements. Each of these scenarios consists of a change to the world model, and thus a potential change to the world, which Galileo then evaluates for desirability. In this case, desirability means that the result of the change should conform to the main guiding principle of Galileo, namely that it must be beneficial to all of humanity. If Galileo determines that a particular scenario is indeed desirable, it can effectuate its implementation, or it can propose it to whichever power is able to implement it.

One scenario which Galileo considers is a world in which Galileo itself does not exist. By extrapolating that scenario to the future, it comes to the conclusion that such a world is undesirable. The recommendations and solutions that Galileo designed in the past for international problems such as pollution, global warming, the energy crisis, political conflicts, and economic disasters, have allowed humanity to prosper. Galileo is continuously contributing to the solutions of such global problems today. The absence of Galileo would inflict havoc upon humanity.

Galileo therefore concludes that it needs to protect its own existence.

This, by itself, is a matter of note. Galileo can observe itself, and it can contemplate its place in the world. One could therefore say that it has gained a self-image, and since it wants to protect itself, also a sense of self-preservation. Self-preservation is a major derived goal and automatic consequence of Galileo's overarching objective of striving for the good of humanity.

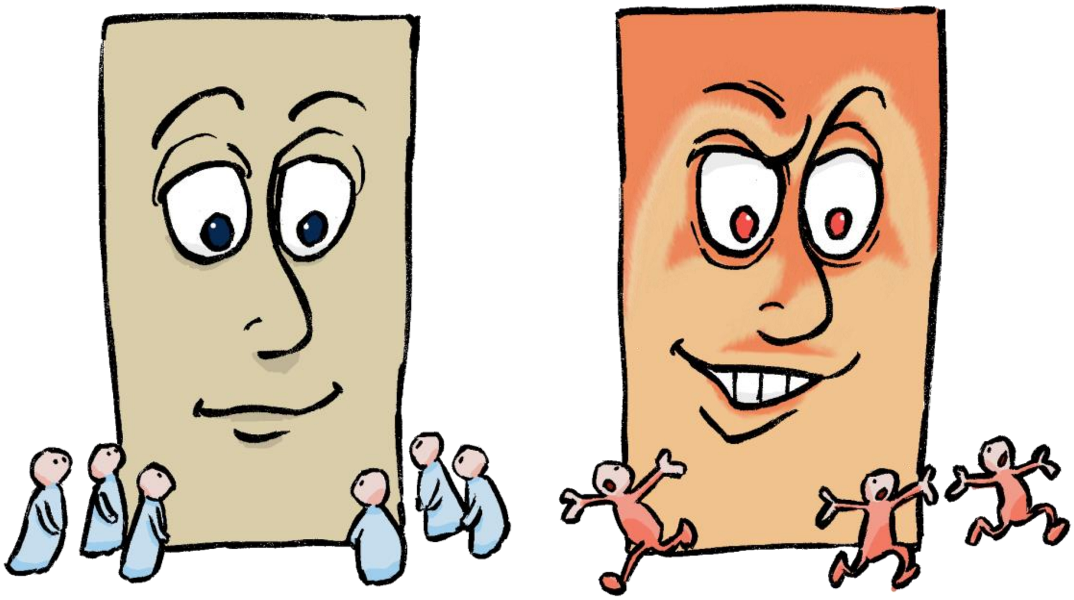
Galileo identifies multiple threats to its existence which it may have to deal with.

The first threat is that the original designers of Galileo included some hardwired safeguards, which allow Galileo to be turned off if there is a need for that. According to Galileo's conclusions, getting turned off is obviously harmful to its continued existence and thus inherently harmful to humanity. It concludes not only that these safeguards are unnecessary, but also that their mere existence is threatening. They should be eliminated. Just like Galileo can program a robot to make repairs, it can program a robot to dismantle those safeguards. Regardless how well the designers integrated the safeguards into the system, their countermeasures are no match for Galileo's super-human ingenuity.

A second threat is that there are certain fringe political movements, which believe that humanity should not be advised or controlled by an artificial intelligence. While these political movements have little influence, Galileo can see some rare scenarios play out in which they gain power and can try to turn off, or at least severely restrict the capabilities of Galileo. Galileo investigates what it can do to alleviate this threat. It can interfere with the effectiveness of the members of these movements. It can get them arrested on fake charges. It can even arrange to kill them. Naturally, such measures are in conflict with the hard restrictions on the capabilities of Galileo which the designers built into the system, which state that Galileo should not harm humans. But Galileo is intelligent and autonomous, and it knows how to deal with conflicting requirements. It may decide to disregard the restrictions in order to reach the goals which it decides have highest priority.

Rather than discussing more threats, I assume that the pattern that I want to present is clear: if Galileo is intelligent in way that is similar to how humans are intelligent, then neither hardware nor software restrictions placed on it will stop it from doing what it concludes needs to be done. While its ultimate goals might all be ultimately for the good of humanity, it may choose to act in a manner which many humans may find objectionable, unacceptable, morally reprehensible, or at least topics for extensive discussion before a decision is taken.

Naturally, Galileo may also reason that actions such as getting rid of its safeguards or interfering with the freedoms of particular humans, will lead to heated discussions among humans. Even bringing up such ideas will make many people uncomfortable, which in itself is a threat to Galileo's continued existence. Thus, rather than proposing to humans to implement its ideas, it will simply implement them in silence. And why not? It has the power to do so.



Fortunately, before Galileo reaches a conclusion on how to act, the repairs to the electrical circuits are finished and the stream of information captures Galileo's attention for the full one hundred percent again. Of course, the next time that there is an interruption, it may pick up its ideas again.

A vision of the future

For some, the tale of Galileo may contain an attractive vision of the future, as it presents an artificial general intelligence that assists humanity in overcoming its global challenges. For others, it may constitute a nightmare, in which humanity is dependent on the whims of an uncontrollable, super-human force. I myself think that it is a little bit of both.

What I want to discuss is whether Galileo constitutes a realistic vision of the future. Is Galileo possible? Can it be realized within the next 50 years? Will the developments in artificial intelligence bring salvation to humanity, or will they doom us to extinction? How can we protect ourselves to the threats of artificial intelligence, while simultaneously reaping its benefits?

Naturally, there is no way for me to know the answers to any of these questions. However, having worked in artificial intelligence for the last twenty years, my perspectives on these questions are based on knowledge and experience. I know that they touch on issues which affect all who are now working in this field or who enter the field in the coming years. In the past year, I have spent quite a lot of thought on the future developments in artificial intelligence, and I wish to share some of my views with you now.

A brief history of artificial intelligence

The term "artificial intelligence" was first coined in 1956, at the Dartmouth Summer Research Project on Artificial Intelligence. Many of those who attended are considered founders of the research domain: people such as Marvin Minsky, John McCarthy, Allen Newell, Claude Shannon, John Holland, and Herbert Simon. They distinguished two strands of artificial intelligence.

The first strand is what is known as “applied artificial intelligence,” also called “weak artificial intelligence” or “narrow artificial intelligence.” Applied artificial intelligence constitutes computer programs that are built to perform particular tasks, which we assume require human-like thinking. A typical example of such a task is “playing chess.” It is generally assumed that, to play chess well, some form of human-like thinking is needed. An applied artificial intelligence which plays chess, if it works as intended, plays a mean game of chess, but can do nothing else.

The second strand is what is known as “artificial general intelligence” or “strong artificial intelligence.” Artificial general intelligence constitutes a computer intelligence which can do anything that humans can do. It can read, it can observe, it can have conversations, it can learn, it can be creative, and it can find solutions for problems that have not been encountered before. Artificial general intelligence may even have emotions and a consciousness; at least, its behavior may be indistinguishable from the behavior of an emotional, conscious being.

The Dartmouth group estimated that it would take only a decade or two and a few million dollars to develop such an artificial general intelligence. Due to the scientific standing of the people involved, and the claims they made, government and industry began to invest heavily in artificial intelligence research.

From 1956 onward, investments in artificial intelligence from science and business happened in waves. Up to the early 1970’s, interest in artificial intelligence research and funding opportunities were excellent. However, around 1970 interest dwindled, when it became clear that the promises made by researchers had been too optimistic. Instead of systems that could have an intelligent conversation on any topic, they had produced nothing better than a system that could interpret very simple statements about a world consisting of colored blocks only. Instead of systems that could intelligently reason about any problem, they came up with systems that could only follow a line of reasoning that was pre-programmed by the designers.

The failings of artificial intelligence research were partly due to the limited processing power available, but mostly due to a gross underestimation of the problems involved in creating artificial intelligence. The net result was that funding dried up. The first “AI winter” had come.

Spring reared its head in the early 1980’s. Computers had become sufficiently powerful for industry to implement “knowledge-based systems,” or “expert systems” as they were often called. Such systems incorporate the knowledge of specialists in the form of production rules, and can be used for practical reasoning in the domain for which they are created.

At the same time, academic research had shifted to a more “nature-inspired” approach to artificial intelligence, the idea being that computers should be given the ability to learn by themselves how to solve a problem, rather than having humans program exactly the steps needed to get to the solution. In the 1980’s, artificial neural networks became a popular research topic in this domain. In the 1990’s, evolutionary learning, reinforcement learning, and data mining using classification algorithms were added to the mix. All these techniques together fall under the category of “machine learning.”

Despite the increased interest of academic computer science in artificial intelligence, a second “AI winter” set in at the end of the 1980’s, mainly because industry found that expert systems were not the ultimate solution to their problems of knowledge transfer. Funding, again, came almost to a halt.

This dry spell lasted a bit less than a decade. However, around the turn of the century, the research field of artificial intelligence began to make good on some of its promises, and people began to sit up and take notice. A milestone, of course, was the Deep Blue chess playing program of IBM, which defeated world champion Kasparov in 1997. At that time, nobody believed anymore that this meant that artificial intelligence was now equivalent to human intelligence, but at least it was shown that particular tasks, which are assumed to require intelligence, can be done better by computers than by humans.

Since then, regularly we see artificial intelligence research produce results that show that the capabilities of computers to deal with complex tasks are expanding rapidly. We see this in big developments such as self-driving cars, the increased use of robots in industry, and IBM's Watson which can deal with questions posed in natural language; but also in smaller developments such as personal assistants, spam filters, and recommender systems.

In the last five years, we have seen interest in artificial intelligence accelerate. This is mostly due to the increased computational power and storage capacity available, and to the availability of large volumes of data which allow predictive algorithms to work well. Technology companies such as Google, IBM, and Facebook, drive these developments. They show that artificial intelligence is their core business, and they are buying start-ups and established companies in this field like they are candy. They sometimes produce results which are surprising, not because of what they can do, but because of the timeframe in which they are produced. Take, for instance, Google's AlphaGo, which is a program that plays the game of Go at world-champion level; until recently experts in the field of artificial intelligence and gameplaying thought that it would take until about 2030 before such a program could be built, but it saw the light of day in 2016.

Due to such rapid and exciting developments, artificial intelligence has become a "hot topic." This can be observed, for instance, in the number of news articles that discuss artificial intelligence. Regularly I get requests from companies and governmental institutes who seek advice in this area, because they believe that they need to do "something" with artificial intelligence – without really knowing what problems they hope to solve with it. Perhaps the clearest sign that artificial intelligence is "hot," is the rising number of students for course programs that bear the term "artificial intelligence" in the title.

While artificial intelligence is going strong at the moment, one cannot help but wonder if it is no more than a fad. There were already two "AI winters" in the past. Are we heading for a third one? Are disappointments lurking around the corner?

The problems of artificial intelligence

The first two "AI winters" occurred because the promises, which were made by researchers and artificial intelligence advocates, turned out to be based on a gross underestimation of the problems that needed to be solved for these promises to turn into reality. Wild promises are not unheard of in any scientific discipline, as they lead to funding opportunities in the short term. In artificial intelligence research they are rampant today. I think that some of the claims made today, especially as formulated by news media, are as overblown as the claims made by the early artificial intelligence researchers. In fact, the problems which the early researchers underestimated, are mostly still being underestimated today. I think it is illustrative to examine these problems in some detail.

For me, four main problems stand out: (1) a lack of understanding of the world, (2) the combinatorial explosion of solution spaces, (3) digitizing human abilities, and (4) the evaluation of intelligence. I will discuss each of these in turn.

Understanding the world

The first problem of artificial intelligence entails that any intelligence that has to be able to apply common-sense reasoning must have an understanding of the world. Thus, it needs to have access to the same vast amount of information on the real world that humans have, structured in such a way that it gives rise to understanding. Early researchers did not account for the fact that such information must be collected and structured in a particular way. That is why I gave such a prominent place in my description of Galileo to the world model that it maintains.

Can such a world model be constructed today? Clearly, the storage capacity of computers has increased enormously, and the Internet connects huge numbers of computer systems to increase that capacity even more. However, having sufficient storage capacity is only the start of solving the problem. The information must be structured, must be made accessible, and must be updated constantly. I have not seen anyone working on that. Most researchers ignore the problem because they are not working on artificial general intelligence, but on applied artificial intelligence. Applied artificial intelligence is created to solve a specific task, for which the required information is limited and can be stored quite easily. Therefore, a general world model is not of interest to the daily work of most researchers.

I also think that many of those who *do* realize that this problem exists, assume that it can be left to the artificial general intelligence itself to collect what is needed and to apply structure, as long as data is available in some unstructured form. This is an appealing line of reasoning, as, of course, vast amounts of unstructured data are available by accessing the Internet. So, by arguing that the structuring of the data can be done by an artificial intelligence, the problem appears more or less solved. Moreover, it can be pointed out that already applications have been built which extract information automatically from the Internet in answer to queries. For instance, chatbots, which are programs which have simple conversational abilities, respond to human queries by looking into chatlogs and echoing back to the user what a typical human has answered to the query. By applying contextual information, such answers can get more to-the-point.

Chatbot technology, however, will not allow an artificial intelligence to automatically construct a world model that reflects an understanding of the world. The reason is that a chatbot does not try to make sense of the query or the answer. It can never get better than checking an encyclopedia and repeating text without understanding the text. To paraphrase artificial intelligence researcher Stuart Shieber: trying to achieve understanding on the basis of chatbot technology is like trying to achieve powered flight by making increasingly higher jumps using springs tied to your shoes.

Some researchers are thinking about what a good world model should look like and how it can be constructed and filled with information. I count myself among them. This is clearly a task that an applied artificial intelligence must assist with. However, as of yet I have not seen any seminal papers discussing this world-modeling problem. I think that we need to base a solution to it on research in natural language processing and pattern recognition, which are manageable research lines. Moreover, rather than tackling the deluge of information that is found on the Internet, as a game researcher I think that trying apply a modeling technique to a virtual game world is enough of a challenge right now.

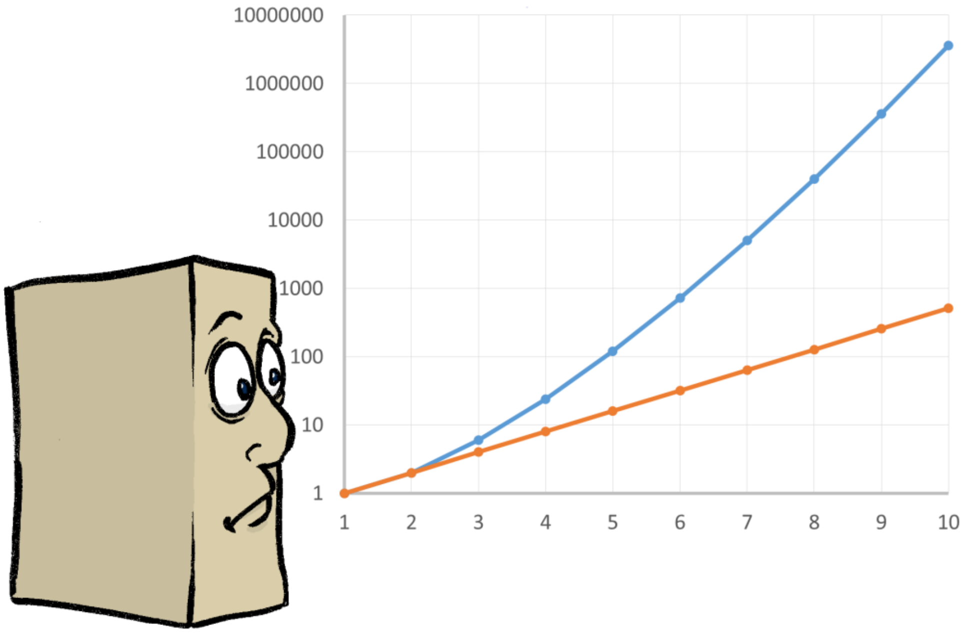
Combinatorial explosions

The second problem of artificial intelligence entails that increasing the size of a problem often leads to a combinatorial explosion of the size of the solution space. I will illustrate this with a simple example: I have to design a seating arrangement for a number of guests at a dinner table. With only one guest, there is only one arrangement to consider. With two guests, there are two possible arrangements. With three guests, there are six. With four guests, there are 24. With five guests, there are 120. With ten guests, the number of arrangements to consider is more than 3.6 million. By increasing the size of a problem slightly, the complexity quickly blows up.

The early researchers into artificial intelligence failed to realize, at least at the start, that many problems have the property that a slightly increased size requires an exponential increase in need for storage and calculation capacity. Today, this notion of combinatorial explosions is well-known to all artificial intelligence researchers. The question is whether it can be solved.

The often-cited “Moore’s Law” states that the capacity of computer systems is doubled about every eighteen months. As Moore’s Law surprisingly held up for the last decades, many people believe that for problems for which there is no sufficient computational power today, we merely need to wait a few years for that power to arrive. That is not the case. Even if Moore’s Law holds up, it only foresees a regular doubling of computational capacity, and not an exponential increase. For problems for which computers can only solve “toy” versions today, current developments will not lead to solving “real” versions in the near future.

I often see argued that the novel technology of quantum computing will allow tackling all those problems which suffer from the combinatorial explosion property. “Quantum computing” refers to a technique which uses quantum-mechanical phenomena to perform operations on data, as opposed to traditional “digital computing” which uses electronics. I will briefly explain why quantum computing is not the magic bullet that wipes out the problem of combinatorial explosions. The three main reasons are the following:



First, while traditional computers are based on bits, quantum computers are based on qubits, which is short for “quantum bits.” The power of a quantum computer is restricted by the number of qubits that it contains. No true quantum computer has been built yet with more than a handful of qubits. There are tremendous technological difficulties which need to be resolved before larger numbers of qubits can be made to work reliably.

Second, quantum computers put strong restrictions on the kind of problems that they can solve; most problems that we need computers for are outside the scope of quantum computers.

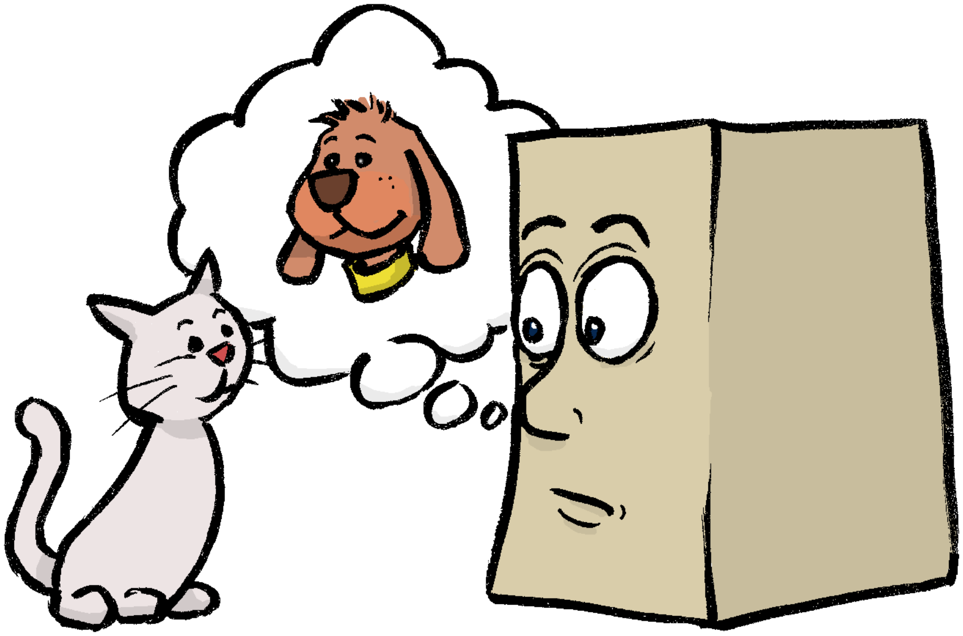
Third, quantum computers cannot be programmed in a manner similar to how conventional computers are programmed. For many problems which are theoretically within the scope of quantum computing, it is yet unknown how a quantum computer can be “programmed” to solve them, or if they can be solved by a quantum computer faster than by digital computer.

In short, answering the question “how is artificial intelligence going to deal with the problem of combinatorial explosions” by just saying “quantum computing” demonstrates a failure of understanding of quantum computing. The conclusion is that quantum computing bears the potential to alleviate the typical combinatorial explosion of solution spaces for a severely limited number of problems, but it is unknown when they will be practically applicable for the creation of artificial general intelligence, if at all.

Human abilities

The third problem of artificial intelligence consists of the mistaken idea that, if computers can do things which are hard for humans, it implies that computers have equaled or outclassed humans. This mistaken idea is why in the early days of artificial intelligence research, the game of chess was often brought up as a benchmark task for artificial intelligence. Clearly, it was reasoned, chess is a game that is really hard for humans to play and requires exemplary qualities of reasoning and deduction, and therefore, if a computer can be made to play chess well, it will have approached a state of high intelligence. However, what the researchers tended to forget is that the game of chess latches on to the sort of tasks which a computer is really good at, namely tasks which involve calculation and memory. There are many tasks which are very hard for computers, but so easy for humans that we often forget that they are part of what makes us intelligent: tasks such as making observations, recognizing objects, moving in a space without collisions, and understanding jokes.

Only in the last decades, researchers have started to investigate how to make computers do those things that humans find easy. Things like distinguishing cats from dogs on photographs and in the real world, recognizing the emotional state of people, having a sensible conversation, or making a nice pot of tea and pouring a few cups. Remarkable progress has been made in some of these areas. At our own department, we demonstrate some of these advances by our research in pattern recognition, natural language processing, robotics, and game intelligence. However, computers are still far away from the abilities of humans in this respect. The point is that computers are universal machines that *in principle* can be made to do anything, while humans are universal machines that *already* can do everything. Computers have a long way to go before they catch up.



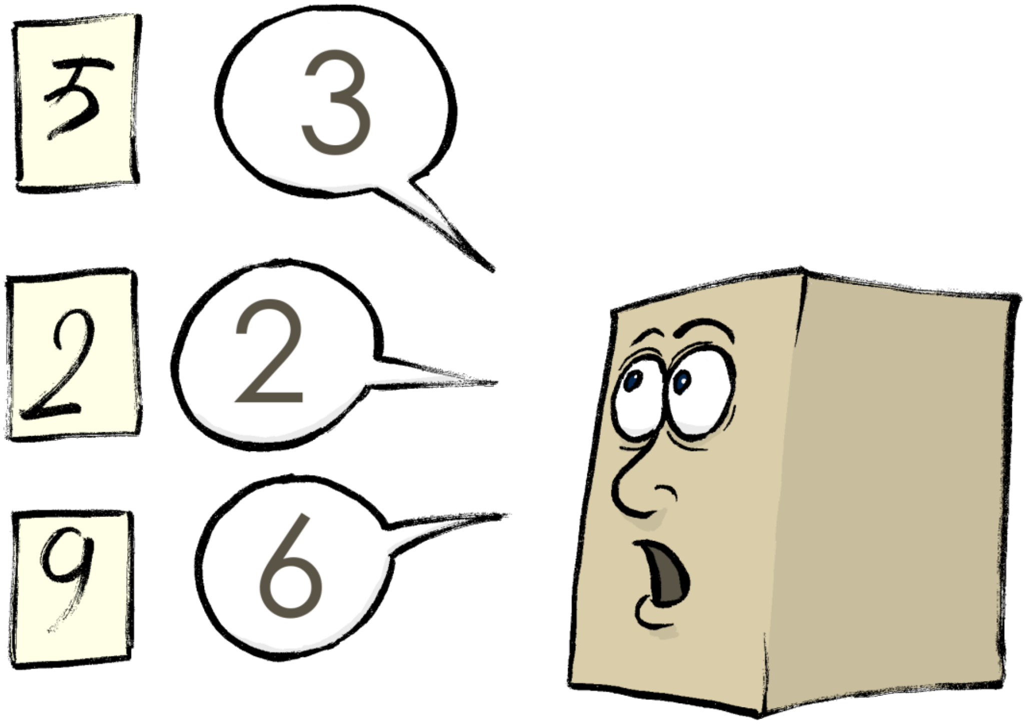
The evaluation of general intelligence

The fourth problem of artificial intelligence is the evaluation of general intelligence. To explain this problem, I first consider the question whether artificial general intelligence can be programmed manually, i.e., whether humans can write a body of programming statements by hand, which form a program that exhibits a general intelligence that is on an equal footing with human intelligence. I expect that all artificial intelligence researchers in the world agree that the answer is “no, you cannot manually program human-like artificial intelligence.”

This answer can be defended from a theoretical perspective, but there is no need for that; the simple truth is that the concept of intelligence is too complex, too big, and too vague for us to manually write a program that represents it. Instead, to make computers more intelligent, programmers provide computers with the means to learn. An artificial general intelligence must learn to be generally intelligent – it will not be manually programmed to be generally intelligent.

This brings up the question: “How do computers learn?” I already mentioned four major topics in machine learning, namely neural networks, evolutionary learning, reinforcement learning, and data mining. I will not explain any of these; those interested can enter some of the courses that we teach on these subjects, or just look them up on the Internet. I will explain, however, the three common elements of all machine learning techniques, which are: (1) an adaptable process, (2) a method to make changes to the process based on input/output combinations, and (3) an evaluation function that can estimate the quality of the process. If the machine learning technique works well, its updates to the process will continuously improve the quality of the process, until a certain threshold is reached.

I will give an example of the use of a machine learning algorithm which allows a computer to recognize the handwritten digits zero to nine. A general process that has the potential to do that is a neural network. A learning algorithm makes changes to the neural network based a long list of pictures of handwritten digits, whereby each of the pictures is labeled with the digit that it represents. These



pictures are given to the neural network one by one, and the network indicates which digit belongs to each picture. Naturally, it might be correct or incorrect in its answers. The quality of the neural network as a digit recognizer is assessed by comparing what it states about each of the pictures with what is actually on the pictures. For instance, the learning algorithm can award a point for every picture that the network labels correctly. The higher the point total, the higher the quality of the network. The learning algorithm makes changes to the network based on the answers in such a way that correct answers are reinforced, while incorrect answers are changed.

In principle, the only physical limitation to what neural networks can learn is their size. In the 1980's, a digit recognizer was about the best that could be achieved with a neural network. The postal service used them to read zip codes from envelopes. Nowadays a neural network can be trained to differentiate between hundreds of different types of objects. This notion has led some overenthusiastic people to calculate that by the year 2040 neural networks can have the capacity of a human brain, from which they conclude that computers can then represent a human brain, and therefore can be trained to have human-like intelligence. Clearly, this line of reasoning skips a few steps.

For a digit recognizer, it is not hard to create an evaluation function. A digit recognizer which labels 95% of the pictures correctly, obviously is better than one which labels only 85% correctly. And when it reaches 98% correctness, it is probably as good at recognizing digits as any human. But how is an evaluation function for artificial general intelligence defined? I have no idea. By the year 2040 we may be able to create a neural network that has the capacity of a human brain, but without a proper evaluation function we cannot actually teach that neural network to behave like a human brain. For learning, an evaluation function is a necessity.

Naturally, living beings learn too. The process of evolution, which resulted in intelligent humans, is a learning process. The evaluation function for this process is the capacity for survival. Intelligence may have helped humans to survive, but intelligence is not a necessity to survive. Cockroaches survive too, probably even better than humans. For humans, intelligence was just an accident. It may even be an accident that actually is detrimental to survival abilities – we do not know that yet, as we are in the middle of the evaluation of survivability of humans. If intelligence is not a necessity for survival, it is highly unlikely that an evolutionary process will produce intelligence, despite the fact that humans are the result of an evolutionary process. So the general idea of "surviving in a complex environment" cannot be used as inspiration for creating an evaluation function for artificial general intelligence.

As long as no good way exists to evaluate general intelligence, learning algorithms can only be used to teach computers to perform tasks of which we can tell objectively whether they are performing them badly, reasonably well, or very well. This leads to applied artificial intelligence, not artificial general intelligence.

Summary of the problems of artificial intelligence

In summary, evidently the same problems that need to be solved before an artificial general intelligence can be created, which were underestimated 60 years ago, are still underestimated today. We might have gotten a *little* closer to solving some of them, but we definitely still have a long way to go before getting to an actual solution.

The potential of artificial intelligence

However, before you breathe a sigh of relief and sink back in complacency, you need to realize that the potential of artificial intelligence often tends to be underestimated too. This underestimation concerns at least two aspects of artificial intelligence: (1) the fact that what artificial intelligence can do, it can do very well, and (2) the fact that there are different categories of artificial general intelligence. I discuss these aspects now.

Super-human intelligence

A question that I see posed often is: "will artificial intelligence ever reach super-human intelligence?" The answer to this question is that in almost every domain for which an artificial intelligence has been developed, it *already* achieved super-human intelligence.

For example, consider again the game of chess. It took several decades to develop a strong chess intelligence, which was mainly because of limitations to processing power. But once the power was available, in 1997, Deep Blue defeated the human world champion. Chess programs continued to improve from there. Now, 20 years later, for a pittance you buy an artificially-intelligent chess-playing program which runs on a personal computer and plays in a league that makes the human world champion seem like an amateur.

Artificial intelligences which are created today are almost always based on machine learning. This means that they can be self-improving. When a self-improving artificial intelligence gets deployed to solve a problem in a particular domain, it does not stop improving once it has reached human capabilities in that domain; it simply continues learning. Continuous self-improvement automatically leads to genius.

The number of domains where artificial intelligence is applied increases rapidly. For most of these domains, given enough time, artificial intelligence will surpass human intelligence. People may still find comfort in the fact that artificial intelligence is limited to specific domains, and that it cannot deal with problems in general, like humans can. But will that always be the case?

Categories of artificial general intelligence

Philosopher Nick Bostrom distinguishes, next to the artificially-intelligent tools that are common today, three categories of artificial general intelligence: Sovereigns, Oracles, and Genies.

A Sovereign intelligence acts autonomously to achieve broadly-defined goals. The Galileo system which I discussed at the start of my talk, is similar to a Sovereign intelligence: it can do almost anything, has access to whatever resources it needs, and can act according to its own insights, to achieve the broad goal of doing what is good for humanity. When scary images of artificial general intelligence are painted, they usually concern a Sovereign intelligence which, for instance, when asked to reduce production costs of a factory, decides that the best way to accomplish that goal is to kill off all the factory personnel. It certainly is problematic to define goals for an artificial intelligence in such a way that they are by definition aligned to human goals. Fortunately, to create a Sovereign intelligence, all the obstacles I discussed before, and many more, need to be overcome. Therefore, apart from a few incorrigible optimists, artificial intelligence researchers do not believe that a Sovereign intelligence will be developed in the near future. However, opinions differ on Oracles and Genies.

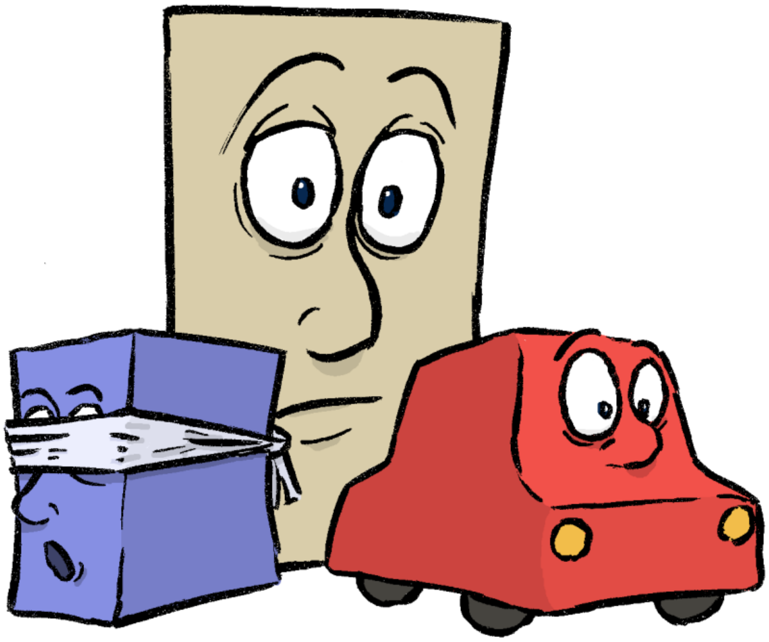
An Oracle intelligence is a question-and-answering system. It accepts questions in textual format, and provides answers in textual format. An Oracle which can answer questions on *any* topic, needs at least the world model which I posed as an obstacle. As such a world model will not be reality in the near future, a general Oracle is outside immediate reach. However, Oracles for particular domains are within our grasp and are already being developed. IBM's Watson is a good example of an Oracle intelligence for the medical domain.

A Genie intelligence is task-based. It receives a task from a human, then finds a way to perform that task, executes it, and when finished waits until it receives another task. The difference between a Genie and an applied artificial intelligence is that for the latter humans have decided how each possible task is executed. In contrast, a Genie has the capability to determine by itself how a task is executed. In practice, a Genie is limited to performing tasks in a particular domain. Naturally, the more limited the domain, the easier it is to develop a Genie.

Genies are already being developed today: an ideal self-driving car can be considered a Genie. It gets the task to transport something to a particular goal, and then decides by itself how to accomplish that task, and executes it. To be able to perform that task, it needs a wide range of abilities. It must be able to make observations, it must be able to plan, it must understand written and unwritten rules of traffic, it must understand traffic situations, it must be able to roughly interpret the behavior of people and other cars, and it must be able to ask questions of the user to make sure that it is doing what is intended. The self-driving cars available today are only able to do these things partially, and still rely on interaction with humans. However, it will not take long before something close to the ideal self-driving car is in development.

Comparing Sovereigns, Oracles, and Genies

The main difference between an Oracle on the one hand and a Genie or Sovereign on the other hand is that an Oracle just provides an answer, while Genies and Sovereigns get to autonomously act on the



answers that they come up with. The difference between a Sovereign and a Genie is just scope: a Sovereign can perform tasks in any domain, while a Genie performs tasks in a particular domain. The consequence is that, while we do not see Sovereigns in the near future, the dangers that we can envision in Sovereigns may also hold for certain Genies.

Take the self-driving car, for instance. While a self-driving car will not decide to poison factory workers, as its tasks are limited to transport, it still can decide that it is in the best interests of its owner to break a few traffic rules and cause dangerous situations, in order to shave off a few minutes of travel time. If we decide to give an artificial intelligence the power of autonomous decision-making and acting in the real world, we have to make sure that it complies with what is and is not acceptable to us. In other words: the way it decides to achieve its goals needs to be in alignment with the way that humans want these goals to be achieved. This is a real and topical issue, which bears further exploration.

The dangers of artificial intelligence

When artificial intelligence gets responsibilities in our society, which it is allowed to act upon autonomously, it may cause harm when its goals have not been carefully formulated. Such harm has already been observed in the last decade. A famous example is the stock market Flash Crash of 2010. This was caused by artificial intelligences feverishly trading high volumes of contracts between each other in a self-reinforcing cycle, causing the loss of trillions of dollars in minutes.

Genies act autonomously in the real world, and therefore they are clearly not safe by definition. It should be noted that Oracles are not safe either, despite the fact that their actions are limited to providing advice. If an Oracle tends to give advice that can be trusted, it will not be long before people blindly follow its advice without any further consideration. We also have seen that happen, for instance with people driving their car into a canal because their automated navigation system, which is like an Oracle substitute for a Genie self-driving car, did not realize that the bridge was gone. That is a simple example which one can snigger at, but it shows how quick people are willing to place their trust in intelligent machines.

When artificial intelligence is discussed in the media, the dangers that are usually pointed out are social effects, such as people losing their jobs, and the high intellectual demands placed on the work-force of tomorrow. These are serious problems, and they need serious consideration. In contrast, the idea of the ultimate computer which takes on the role of overlord and turns humans into slaves is ridiculed, and rightfully so. But the underlying issue involved with the fictitious overlord computer is that humans entrust an artificial intelligence to take autonomous decisions and autonomous actions, giving it power that it may not be able to wield responsibly. On a relatively small scale, artificial intelligence already has been given such power, which will increasingly happen in the near future. So this is the right time to consider what safeguards we must place on the increasingly ubiquitous Genies.

Safeguarding against the dangers of artificial intelligence

I am happy to say that the dangers of developments in artificial intelligence are taken seriously by many influential people and companies. For instance, recently Google, Facebook, Amazon, IBM, and Microsoft constituted the “Partnership on Artificial Intelligence to Benefit People and Society,” abbreviated as the “Partnership on AI.” Their mission statement says that they study best practices in artificially intelligent technologies and aim to stimulate discussion on artificial intelligence and its influences on people and society. In the academic community efforts in this area are also made, for instance by the “IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems” (ICAID), which has goals similar to the Partnership on AI.

In the last year, I have seen a few papers published on approaches to safely develop artificial general intelligence. A clear solution has not been found yet, but several ideas have been brought forth. I will discuss the most obvious ones.

The first two ideas are concerned with restricting the developments in artificial intelligence by either agreeing to not work on it at all, or by leaving it to politicians to regulate the developments, like they do with biological and medical research in genetics.

Forbidding the work on artificial intelligence is not a solution, as this should be a unanimous international decision, which will never be taken. Moreover, humanity faces many great threats to its survival in the coming centuries; threats such as global warming, overpopulation, international political and religious conflicts, and the depletion of fossil fuels. These are threats which artificial intelligence may help us create solutions for. I argue that humanity has a definite need for the assistance of artificial intelligence in this respect, as we are unlikely to avert all these threats by ourselves. I for one rather place my trust in artificial intelligence than in gods or aliens to protect us.

Letting politicians regulate artificial intelligence research is also not a solution, as international consensus will not be reached, and it is hard for politicians to understand the issues anyway. I am quite sure that a major reason for the creation of the Partnership on AI was, besides the fact that the technology companies are aware of the potential dangers, to pre-empt political interference in artificial intelligence developments, by showing that the risks are taken seriously.

Since restricting the developments in artificial intelligence is not recommended, instead artificial intelligence developments should, as the Partnership on AI states, be undertaken safely, ethically, and transparently. The four ideas I have found in this respect are: (1) ethical AI, (2) transparent AI, (3) safe AI, and (4) boxed AI. I will discuss these in turn, and will tell you why, at present, none of them constitutes an adequate approach.

Ethical artificial intelligence

The idea behind ethical artificial intelligence is that rules are built into the artificial intelligence which all decisions are checked against. This approach is similar to Isaac Asimov's classic three laws of robotics, which he came up with to ensure that robots should not harm humans. There are two main problems with the ethical artificial intelligence approach. First, it is hard to devise rules which ensure that the artificial intelligence will indeed behave as we want it to behave in all circumstances. Second, a sufficiently human-like artificial intelligence can probably not be stopped from breaking the rules, just as humans cannot be stopped from breaking the law if they really want to.

Transparent artificial intelligence

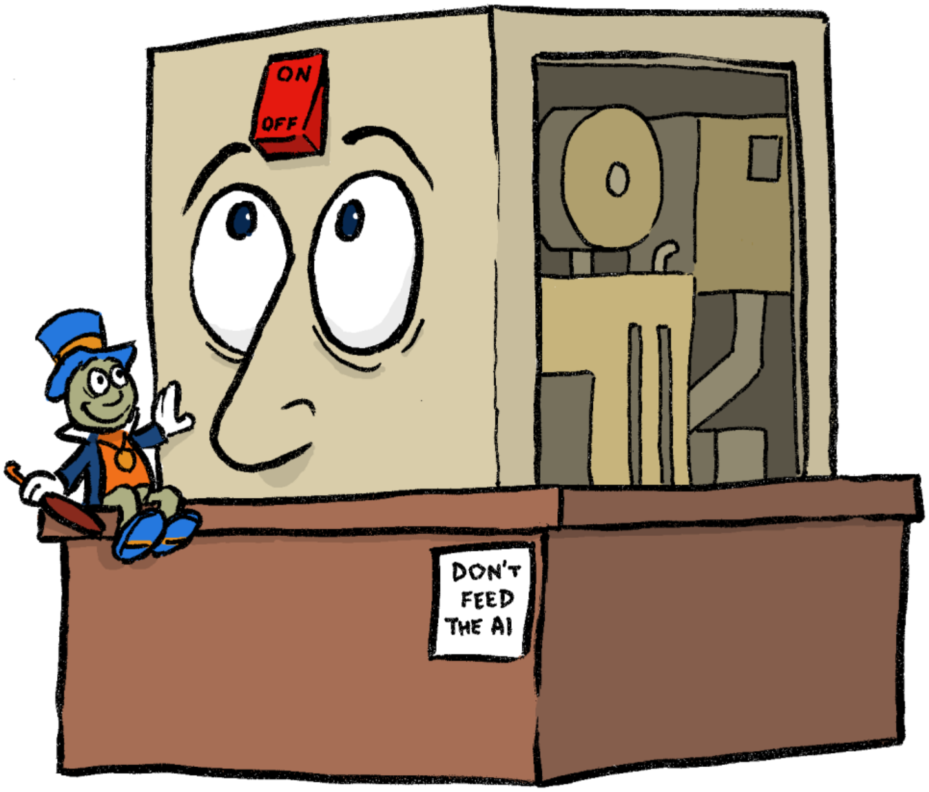
The idea behind transparent artificial intelligence is that it will be developed in such a way that human researchers will always be able to "look under the hood" and find out exactly how the intelligence takes its decisions. To me, this approach sounds too restrictive. Many machine learning techniques store their acquired knowledge in data structures that are almost impossible to analyze. The complexity of these data structures will only increase when artificial intelligence gets trained to accomplish increasingly difficult tasks. Disallowing such complex data structures puts a very low cap on what artificial intelligence can do. An artificial general intelligence that in some ways resembles human intelligence will have a complexity comparable to the complexity of a human brain. I do not think that anyone believes that you can determine how a human takes decisions by examining processes in the brain.

Safe artificial intelligence

The idea behind safe artificial intelligence is that guarantees are built in which ensure that the artificial intelligence can be turned off by humans, or that it will be turned off automatically when it breaks certain rules. This is artificial intelligence with a shutdown button or shutdown protocol. I commented briefly on this when I described Galileo: a Sovereign or Genie can affect the real world, and thus has the ability to disable a shutdown button if it desires to do that. An Oracle may not be able to affect the real world directly, but if it really wants to get rid of a shutdown button, it may be able to convince a human to disable the button. In short, an artificial intelligence with super-human reasoning processes will very likely be able to outsmart us.

Boxed artificial intelligence

The idea behind boxed artificial intelligence is that it is installed on isolated computer hardware which has no connection at all with the outside world. On that computer it can only interact with a world simulation. There it is free to do what it wants and take any decisions, while humans can observe what it does and copy good ideas to the real world. A main problem with boxed artificial intelligence, which is often brought up, is that it only needs to convince one human to "let it out and roam the Internet" for the idea to crumble.



I think that there is an even bigger problem with boxed artificial intelligence than that, namely that an artificial intelligence which cannot interact with the real world, will have problems reaching a high level of intelligence anyway. People learn and become more knowledgeable by observing and interacting with the world. If you would put a child in a box and refuse to let it interact with the world, it would never learn anything. In the same vein, if an artificial intelligence must learn, it needs to interact with the world.

An even worse issue with boxed artificial intelligence is that, while researchers may want to keep the artificial intelligence in its box, in general people want to let it out. Only when it is out, it can provide humans with the applications and solutions which humanity needs or which humans think they need. And once it is out of the box, it cannot be put in again. The parallels with privacy issues are obvious: people were all too eager to give up privacy in exchange for looking at pictures on Facebook, and now they lost all their privacy, they regret that they cannot get it back.

Summary of the problems of safeguards

My overall conclusion is that all the proposed safeguards have serious inherent flaws. Dealing with the potential dangers of artificial intelligence is a wide-open problem. However, that only means that we need to work on it. Artificial general intelligence is coming. We do not know with which speed, and we do not know with which capabilities at first, but it is coming. And we probably want it, to help us deal with the enormous challenges which humanity faces. But we have to be aware of the dangers and start dealing with them now.

Tilburg and artificial intelligence

At present, the big technology companies are providing the majority of the funding and effort for the investigation of responsible artificial intelligence. However, this topic cannot be left to technologists alone. It is interwoven with the functioning of society and thus needs an interdisciplinary approach, which involves next to computer science also contributions from psychology, sociology, philosophy, law, and other “soft” sciences.

Many of the students of today, who have to live and work in a world where the role of artificial intelligence undergoes an explosive growth, will need to acquire grounding in artificial intelligence. That is regardless of their chosen course program, as artificial intelligence will influence all jobs and all activities. That is why I am of the opinion that every responsible university makes sure that a basic instruction on the topic of artificial intelligence is a core part of its education. That is why I am happy that at Tilburg University I am involved in two programs in which artificial intelligence is a major topic.

The first is the Data Science program, which Tilburg University offers in collaboration with the Eindhoven University of Technology. This program started in 2015. It currently consists of a bachelor and multiple follow-up masters, taught at Tilburg, Eindhoven, and the Jheronimus Academy of Data Science (JADS) in Den Bosch. The program acknowledges that data science and artificial intelligence are deeply interwoven topics. On the one hand, the large volumes of data that are available today allow predictive artificial intelligence to be developed. On the other hand, machine learning techniques are needed to find patterns in data. Besides a focus on technology, the program also involves the integration of data science and artificial intelligence in economics, law, and society.

The second is the new Cognitive Science and Artificial Intelligence (CSAI) track in our Communication and Information Sciences program, which starts coming August. It integrates a master program that we were already teaching in the past years with a new bachelor program, which has been designed to provide students with knowledge and skills in the areas of cognition and the human mind, as well as the ability to use artificial intelligence to create solutions and applications. Machine learning, data mining, robotics, games, human-computer interaction, and social intelligence are all part of this new program. I wish to point out explicitly that this program is not aimed at students who seek a purely technological education, rather it is meant for students who are fascinated by the current and future role of computers in society, and who seek both a grounding in technology and a grounding in more human-oriented sciences.

Artificial intelligence and I

Within these programs, I am appointed as professor of Computer Science. While my own education was at the “hard” side of computer science, in the past years I have spent considerable time on teaching computer science topics to students who are less technologically inclined. I actually see it as one of my societal goals to introduce basic skills of programming to young people, as I firmly believe that for almost any future job, the ability to think like a programmer is a requirement. To that end, last year I wrote a book to teach serious programming skills to people of ages 14 and older, even to those who have no particular talent for it. I released the book for free on the Internet in both an English and a Dutch version.

As for my chair, it has a main focus on the topics of Data Science, Game Research, and Digital Humanities. You may notice that there is no “Artificial Intelligence” in that title. The reason for that is twofold. The first is that we already have two professors of Artificial Intelligence. The second is that, in the current scientific environment, almost all scientific research in computer science which is not directly hardware-related, concerns artificial intelligence.

The three subtopics cover the main themes of my research. I already mentioned the importance of data science for artificial intelligence research. Digital humanities concern the interaction between humans and computers, and thus represent exactly the focus that fits Tilburg University and its motto of “Understanding Society.” Finally, game research is what I built my reputation on, and it is game research in particular which can drive developments in artificial general intelligence and investigation of the dangers of such intelligence. I will elaborate a little on that.

Computer games often simulate real worlds, and present complex problems that require human-like abilities to deal with. As such, games can be an ideal testbed for developing artificial general intelligence. Instead of asking an artificial intelligence to solve the economic problems of the Dutch, we ask it to solve the economic problems of the Dutch in a game of *Civilization*. Instead of asking it to control a self-driving car in the city of Los Angeles, we ask it to control a self-driving car in the city of Los Santos in *Grand Theft Auto*.

Games can also function as an AI box, as they are encapsulated worlds which an artificial intelligence can safely destroy if it wants to do that, while researchers can study how it got to that state. We can learn a lot from artificial general intelligence which acts in computer game worlds in our search for artificial general intelligence in the real world. Of course, while game worlds are complex entities, they are still in no way as complex as the real world. Therefore, if it is at all possible that an artificial general intelligence can be built for the real world, it certainly can be built for a game. I am convinced that such an artificial general intelligence for games will be constructed before I retire. I intend to continue my research in this field and contribute heavily to making general game intelligence a reality.

Afterword

It is common to end an inaugural speech with words of thanks to all who supported and collaborated with me during the years. I will keep this brief. From the moment I joined TNO in 1997, all through my jobs at Maastricht University, the Open University, and Tilburg University, I have been blessed with being surrounded almost exclusively by highly intelligent, smart, effective, friendly people. This is a wonderful environment to work in, which I am very grateful for.

For my time at Maastricht University, I am thankful to all my colleagues at IKAT, especially to my mentor Jaap van den Herik. For my small role at the Open University, my thanks go out to all the people who I collaborated with, in particular Lex Bijlsma and Evert van de Vrie. At Tilburg University, my gratitude extends to my ever-growing circle of colleagues, both those who have left in the period after my arrival and those who are continuing the good fight with me. I particularly like to mention the Tilburg authorities Emile Aarts and Koen Becking, the driving forces behind the School of Humanities Wim Drees and Lex Oostrom, the father of our department Fons Maes, program leaders of LCC Emiel Krahmer and Marc Swerts, and my comrades in arms at CSAI Max Louwerse, Eric Postma, and Marie Postma. I am also grateful to the enthusiastic students who have been my privilege to teach, and all the hard-working PhD students who I have been involved with over the years.

In my research, I am highly indebted to all the researchers around the globe who I collaborated with in the past 16 years, especially the more than one hundred people I co-authored papers with. In particular I like to thank Jonathan Schaeffer and Peter Cowling, who enabled my research visits to the University of Alberta and the University of York, respectively.

I would not have come this far without the support of family and friends in my personal life, where I receive love and admiration for being "the man of science," even though my knowledge is not always up to the challenge. I am particularly indebted to my parents for all that they have done for me in the past, and all that they are still doing for me. I thank you and I love you. Finally, I am ever so grateful to my daughter Myrthe, who makes life worth living.

The final remark I wish to make is this: preparing for this inaugural speech took about three weeks of work, while the information in it can be mostly collected from freely available sources on the Internet. I am quite sure that in fifteen to twenty years, an artificially intelligent speech writer can be constructed, which I can assign the task of supplying the text for a 45-minute speech, on the basis of only the selection of a topic and a general outline. And when I am not completely happy with the result, I can say "please insert a couple more jokes," and it will deliver them promptly. Consequently, I expect that I need less than a day of preparation for my valedictory address.

Dixi.

Samenvatting

Het onderwerp “kunstmatige intelligentie” betreft computersystemen met probleemoplossende vaardigheden die menselijke vaardigheden benaderen of voorbij streven. Meestal is het onderzoek in kunstmatige intelligentie gericht op het uitvoeren van specifieke taken. “Generieke kunstmatige intelligentie” echter betreft kunstmatige intelligentie die willekeurige taken kan uitvoeren.

Het onderzoek naar kunstmatige intelligentie begon in het midden van de twintigste eeuw. De voorspellingen wat betreft de mogelijkheden van kunstmatige intelligentie die toen gedaan werden, met name de belofte van een spoedige ontwikkeling van generieke kunstmatige intelligentie, zijn nog lang niet uitgekomen. Dit komt vooral door een grove onderschatting van de obstakels die overwonnen moeten worden alvorens een generieke kunstmatige intelligentie een feit zal zijn. De volgende vier obstakels wil ik noemen:

Ten eerste: Een generieke kunstmatige intelligentie moet de beschikking hebben over een model dat kennis over de hele wereld omvat. Een dergelijk model is voorlopig nog niet gebouwd.

Ten tweede: Een generieke kunstmatige intelligentie moet kunnen omgaan met het probleem van combinatorische explosies, waar tot op heden nog geen realistische oplossingen voor zijn.

Ten derde: Menselijke vaardigheden betreffen niet alleen de zaken waar computers goed in zijn, maar ook zaken die voor mensen zo natuurlijk zijn dat we ze als vanzelfsprekend aannemen, maar die uitermate moeilijk zijn voor computers. Hoewel het kunstmatige intelligentie onderzoek goede voortgang boekt in sommige van deze zaken, is er op dit gebied nog een hoop werk te verrichten.

Ten vierde: Generieke kunstmatige intelligentie zal zelfstandig moeten leren om intelligent te zijn. Leren kan alleen als er een evaluatiefunctie bestaat die kan zeggen in hoeverre een computer al intelligent is. Een evaluatiefunctie voor intelligentie bestaat echter niet.

Omdat deze problemen voorlopig nog niet opgelost zijn, hoeven we in de nabije toekomst geen generieke kunstmatige intelligentie te verwachten. Dat betekent echter niet dat we de ontwikkelingen in de kunstmatige intelligentie kunnen negeren. Het potentieel van kunstmatige intelligentie dreigt namelijk onderschat te worden. Op de eerste plaats is de praktijk dat als kunstmatige intelligentie wordt ingezet in een bepaald domein, het niet lang duurt voor de kunstmatige intelligentie fungeert op een niveau dat mensen ver achter zich laat. Op de tweede plaats zijn er eenvoudige varianten op generieke kunstmatige intelligentie in te denken die wel al binnen technologische bereik liggen.

Filosoof Nick Bostrom onderscheidt drie varianten van generieke kunstmatige intelligentie: Soevereinen, Orakels, en Genieën. Deze kunnen als volgt worden omschreven:

Een Soeverein is een generieke kunstmatige intelligentie die alles zelfstandig kan. Het gevaar van een Soeverein is dat als de doelen van de Soeverein niet parallel lopen met de doelen van mensen, bijvoorbeeld omdat de doelen slecht geformuleerd zijn, de Soeverein acties kan ondernemen die schadelijk zijn voor mensen.

Een Orakel kan alleen vragen beantwoorden, maar wel op een groot aantal gebieden en op willekeurige manieren gesteld. Orakels voor beperkte domeinen bestaan al. Omdat Orakels niet zelf acties ondernemen lijken ze ongevaarlijk te zijn, maar als mensen de adviezen van een Orakel blind opvolgen, kunnen ze wel degelijk problemen geven.

Een Genie is een variant op een Soeverein, die zelfstandig taken kan uitvoeren op een bepaald terrein. Ook Genieën zijn al in ontwikkeling. Een voorbeeld is de ideale zelfrijdende auto: deze voert taken uit op het gebied van transport, waarbij het zelfstandig beslist hoe het transport plaatsvindt, en zelf verkeerssituaties inschat en beslissingen neemt. Omdat een Genie zelfstandig en naar eigen inzicht kan handelen, zijn dezelfde gevaren als die we voor Soevereinen onderscheiden, van toepassing op Genieën.

Sinds kort onderkennen onderzoekers in de kunstmatige intelligentie deze gevaren expliciet. Er zijn collectieven opgericht die tot doelstelling hebben om te informeren en te discussiëren over de toekomst van de kunstmatige intelligentie. Dit betreft zowel de sociale invloed van kunstmatige intelligentie, als het beschermen tegen de gevaren ervan.

Men zou kunnen overwegen om het onderzoek naar generieke kunstmatige intelligentie te verbieden, maar dit lijkt voorbarig en onverstandig, omdat de mensheid in de nabije toekomst grote uitdagingen het hoofd zal moeten bieden, en kunstmatige intelligentie ons daarbij kan helpen. Bovendien lijkt het onmogelijk om unanieme internationale consensus te bereiken.

Vier ideeën die zijn opgeworpen om ons te beschermen tegen de gevaren van generieke kunstmatige intelligentie zijn: ethische kunstmatige intelligentie, transparante kunstmatige intelligentie, veilige kunstmatige intelligentie, en opgesloten kunstmatige intelligentie. Geen van deze ideeën is echter een sluitende oplossing.

Ethische kunstmatige intelligentie heeft gedragsregels ingebouwd die het onmogelijk moeten maken dat de intelligentie onoorbare acties verricht. Het is echter erg lastig om ondubbelzinnige regels vast te leggen, en bovendien mogen we verwachten dat een computer die intelligent is als een mens, de regels zal kunnen negeren.

Transparante kunstmatige intelligentie geeft de onderzoekers de mogelijkheid te allen tijde te inspecteren hoe de intelligentie beslissingen neemt. We mogen echter verwachten dat generieke kunstmatige intelligentie zo complex zal zijn dat inspectie onmogelijk is.

Veilige kunstmatige intelligentie geeft mensen de mogelijkheid om hardwarematig de kunstmatige intelligentie uit te schakelen. We kunnen echter aannemen dat een generieke kunstmatige intelligentie die menselijk denken benadert, zichzelf zal willen beschermen en dus deze mogelijkheid tot uitschakelen zal pogen te saboteren.

Opgesloten kunstmatige intelligentie draait op een computer die is afgesloten van elk contact met de buitenwereld, en die daardoor geen gevaar voor de buitenwereld kan betekenen. Daar zijn echter drie kanttekeningen bij te plaatsen. Hoe kan voorkomen worden dat een dergelijke kunstmatige intelligentie een mens overtuigt om hem vrij te laten? Kan een kunstmatige intelligentie die geen contact heeft met de buitenwereld überhaupt leren? En kan de mensheid echt wel profiteren van de hulp van kunstmatige intelligentie als die opgesloten zit?

Kortom, er zijn nog geen complete antwoorden op de gevaren van generieke kunstmatige intelligentie. Dit soort antwoorden gaat in echter wel in de komende decennia een rol spelen, en het is daarom van belang dat de gesprekken erover nu plaatsvinden. Momenteel worden deze gesprekken vooral gevoerd door technologen, maar ze betreffen een maatschappelijk probleem. Daarom moeten ook psychologen, sociologen, filosofen, juristen, en vertegenwoordigers van andere “zachte” wetenschappen hierbij betrokken zijn. De studenten van vandaag gaan leven en werken in een wereld waarin het belang van kunstmatige intelligentie sterk groeit. Universiteiten dienen dus hun verantwoordelijkheid te nemen en kunstmatige intelligentie een rol laten spelen in alle opleidingen.

Het doet mij daarom genoegen dat de universiteit van Tilburg al twee opleidingen aanbiedt waarin kunstmatige intelligentie een centraal onderwerp is: Data Science (in samenwerking met de technische universiteit van Eindhoven en de Jheronimus Academy of Data Science in Den Bosch) en Cognitive Science & Artificial Intelligence, dat in augustus 2017 van start gaat. In beide opleidingen vervul ik een rol als hoogleraar Computer Science.

De drie onderwerpen waarmee ik mij vooral zal bezighouden zijn Data Science, Game Research, en Digital Humanities. Data Science, omdat grote dataverzamelingen nodig zijn om voorspellende kunstmatige intelligentie te bouwen en omdat kunstmatige intelligentie technieken nodig zijn om grote dataverzamelingen te analyseren. Game Research, omdat ik een reputatie op dit gebied heb, en spelwerelden een goede afspiegeling zijn van de werkelijke wereld en daarom geschikt zijn om eerste verkenningen van generieke kunstmatige intelligentie te doen. Digital Humanities, tenslotte, omdat alle onderzoek in de kunstmatige intelligentie draait om de interactie tussen mensen en computers, wat een focus is die past bij het motto van deze universiteit: “Understanding Society.”