A Study on the Ethics of AI

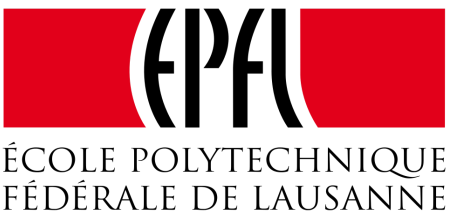
From Superintelligence to Artificial Consciousness

HUM-412c - Ethics of Engineering,
Final Report

20/05/2018

Group I
Andrea Mussati, Felix Nilius, Diana Petrescu, Xavier Willemin

Supervising Professors: Hugues Poltier, Gaia Barazzetti



Index

1. Introduction

As future engineers in Computer Science and Microengineering, we will inevitably be confronted with, or even participate in, the development of Artificial Intelligence (AI).
While AI is in an embryonic phase at the moment, further advancements could lead to the creation of a Super Intelligence (SI), whose computing and task-solving capabilities - superior to those of any human - may cause radical changes and shifts in our society and way of living, to the point that the use and development of SI deserves an ethical analysis.
AI and SI can with no qualms be considered as mere tools, as another instrument built by humans for humans. Yet moral quandaries rise when, either purposefully or by accident, a further step is taken and we beget Artificial Conscience (AC) and Artificial Consciousness (ACn): when our invention, our creation can be considered as a moral agent and a moral subject.

1.1 Abstract

Definitions are given for superintelligence, artificial conscience, and artificial consciousness. For each of those, the related ethical questions are explored.
For SI, those relate mostly to SI usage and regulation, in particular in relation to who controls it; for AC, they are tied to how we should define and build AI morality; while for ACn, they are mostly related to the treatment of ACn.
Various scenarios on the possible evolution of AI and what it would mean for human society are then proposed in order to suggest a preliminary judgement on whether and how it is worth to develop SI, AC, and ACn.
Finally, we discuss the costs and benefits of conducting research and development on AI, as well as pitfalls to watch out for when interacting with such systems.

1.2 Changes since 1st semester report

The present document is an improvement of the 1st semester deliverable. We first conducted the interviews of experts mentioned at the end of the 1st semester report and used the gathered information to better formalize our definitions and add weight to our opinions. We also took into account the feedback received orally during the defense in December 2017 as well as the written comments on the written deliverable. Finally, we added an ethical evaluation and refined it based on the comments received based on the pre-final deliverable. Of course, the section on the tentative methodology for future work has been removed.

1.3 Document outline

The document is organized as follows:

1. *Section 2* enunciates the central ethical issue. Additionally, definitions, abbreviations and secondary ethical questions are listed for each of the types of AI we will treat - Super Intelligence, Artificial Conscience, and Artificial Consciousness.
2. *Section 3* provides generic background information on the state of the art in AI.
3. *Sections 4 through 6* treat more in detail the ethical questions related to these three types of AI. An actions/actors/effects analysis is performed for each AI type, when appropriate.
4. *Section 7* gives a first overall ethical assessment for the central ethical issue, through a possible scenarios-based approach.
5. *Section 8* discusses if the research and development on AI should be pursued and if yes, under which conditions.

2. Ethical Issue and Definitions

2.1 The ethical question

Over the course of our research, we intend to study the ethical questions and possible effects of different forms of advanced artificial intelligence.

The general ethical questions can be stated as follows: **Should we pursue development of artificial intelligence? If so, or if artificial development is unavoidable, in which way and which contexts should it be pursued? How should it be regulated?**

Artificial Intelligence can take different forms, and be used in countless applications: it is a multifaceted concept and, likewise, the enunciated central ethical question has different facets to it, hinted at in section 2.2, further developed in chapters 4 through 7, while our conclusions are presented in chapter 8.

2.2 The definitions of Intelligence and AI

In the early stages of AI development that we are facing, it may be hard to tell what is true artificial intelligence rather than for example a scripted behaviour. The fact that, for ease of communication or marketing purposes, many scripted behaviours or simple algorithms are currently called “AIs” (from the 70’s first “communication AIs” to modern “game AIs”, for instance) doesn’t help the public discourse. But even beyond those basic cases, is a simple machine learning algorithm an “AI”? Is the core of a self-driving car an “AI”? According to

Prof. Boi Faltings[1], it isn't possible to define exactly what AI is and there isn't any consensus in this sense between researchers. Therefore, it is imperative to take a step back and first define what intelligence and intelligent programs are.

2.2.1 A definition for intelligence

Indeed, we can find a similar issue in the more familiar fields of biology, psychology and philosophy. From the simple amoeba behaviours - that most would define "mindless" - to the elaborate human ones, going through all the degrees of intelligence that nature displays, finding the boundary between intelligence and non-intelligence is a similar, and similarly hard, exercise.

Various definitions mention elements of logic, creativity, problem solving, planning, or learning. Many point towards the latter, underlining the ability to selectively extract information from experienced situations, integrating it as knowledge, remembering and applying it to future situations where it may be applicable.

The definition for intelligence seems to be destined to remain hazy: not only because there exist varying degrees of it and we cannot be sure where to place the boundary between "non-intelligent" and "intelligent", but more so because it represents a collection of complex, interlinked behaviours and capacities that take many different forms.

2.2.2 The Chinese room and two definitions for artificial intelligence

Imagine that AI research would advance to the point where a digital computer running a certain program would be able to make intelligent conversation in Chinese such that a native Chinese person couldn't tell whether it was a machine - in other words, it would pass the Turing test.
Now imagine an Englishman, who does not know a single word of Chinese, placed in a closed room with the complete set of instructions that constitutes the aforementioned program, and enough paper, pencils, erasers, and time. Such a man would be able, following the instruction set, to simulate Chinese conversation - while not understanding any of it - just as well as the computer did. In fact, John Searle, the author of this mind experiment[2], argued that there would be no essential difference between the two. Both run a program step by step, producing a behaviour that would seem to demonstrate intelligence.

Searle argues that, like the man in the Chinese room does not understand any of the conversation, a digital computer running a program, no matter how advanced, would not *understand* the conversation; and that without *understanding* it cannot be defined as *thinking*.
He distinguishes two types of AI: Strong AI, that would have a mind in the same exact sense humans have a mind, and Weak AI, that would merely simulate the behaviour of an intelligent (and conscious) mind. Searle argues in his paper - which was destined to become

[1] Prof. Boi Faltings. Head of the Artificial Intelligence Laboratory (LIA), EPFL.
[2] SEARLE, John, 1980. Mind, Brains, and Programs. Cambridge University.

one of the most influential in the field of AI research - that only the latter is possible in a digital computer.

It would take a radical shift in how we build AIs in order to reach Strong AI. As such, for most of this paper we will focus on Weak AI and what it would mean for humans, actually intelligent and conscious beings. We still discuss the ethical implications of reaching a Strong AI in section 6.

2.2.3 A different way to look at intelligence: intelligent behaviour

We interviewed a few experts on the topic of AI. One of them, Aude Billard[3], quoted John Searle in agreement, arguing that due the fact that "intelligence" is quite a vague and ill-defined concept, it would be more useful to talk of "intelligent behaviours" - and the capacity of learning those behaviours - rather than artificial intelligence.

"Artificial Intelligence": an artificial mind. The words themselves seem to forsake the acts, to forgo the Body that must go with the Mind, to artificially create a divide among them. Prof. Billard argues that it makes no sense to talk of one without the other (concept of embodiment); that focusing on the action, the behaviour, would shed more light on what an artificial intelligence must be - an Intelligent Artificial Agent.
She illustrated her words with an example. Our way of "understanding" what a key is is linked with recognizing how it's used. Its representation in our mind *is* the act of putting it in a lock and opening doors, the muscle memory of grabbing and turning it, the sensations that come with entering a new space. A disembodied mind, an AI that is just an AI, would have none of that. All it could do, for instance, would be "recognizing" images of keys in the same way the google search algorithm comes up with images of keys when one looks for them.

2.2.4 A definition for intelligent programs

According to Prof. Faltings, an intelligent program is a program that makes decisions by itself and by optimization. Every decision is made according to the circumstances of the moment (as opposed to a "classical" program where everything is predefined). The intelligent program is always supposed to make the right decisions based on concrete conditions. Such decisions can be either made in real time or prepared in advance as part of a learning process that anticipates and analyses as exhaustively as possible the circumstances.

2.3 Beyond current AI: the language of our research

Our research is oriented towards the future of AI. As such, we shall examine the forms it will take. To do so, we'll need to name them (superintelligence, artificial conscience and artificial consciousness) and to define them, as we do in the next three sections.

[3] Prof. Aude Billard, Head of the Learning Algorithms and Systems Laboratory (LASA), EPFL.

2.3.1 Superintelligence

A superintelligence (SI) is a universal artificial intelligence outperforming humans in a vast majority of tasks. By universal, we mean that the artificial intelligence (AI) is able to solve any problem as long as it is possible to find relevant information. Therefore, the current AI, even if outperforming humans for some specific tasks do not fulfill the requirements because one can't ask it any arbitrary task. In fact, the philosopher Nick Bostrom points out that the omni-competence of the SI should also include areas that were for a long time considered as exclusively human (such as scientific creativity, general wisdom and social skills).[4]

We reckon SI will be a form of AI that could have a large-scale impact on human society and way of living. One big ethical issue is the consequence of the use of SI by a minority (misguided governments, rogue state, big companies and/or rich persons) for their own interest in detriment of the common interest. Another one is the consequence of a badly formulated request, which might be solved optimally in detriment to other (ethical) considerations. However, it is a fantastic tool that might solve many of our problems if well used.

2.3.2 Artificial conscience

A superintelligence might be a great tool, but is very dangerous to produce because of the potentially dramatic consequences misuses might induce. A solution would be to transform the SI into a moral agent, having its own representation of moral and ethics. This moral agent is called an Artificial Conscience (AC). In its most simplistic way, this might just be done by adding some extra terms to the cost function of the optimization in order to penalize actions that harm other people. In the most advanced way, the AC would be able to calculate reliably all the good and bad some action might have, in the present and in the whole future, and estimate if this action does or does not increase the happiness in the world.

The main ethical question is how to define an ethic for the AC: it is already very difficult for a few humans to agree on the most ethical solution for very simple problems. Agreeing on general ethical principles that are still applicable is almost impossible. So how to manage the even greater challenge of implementing it in an AC? What ethical strategies should we adopt? Should we seek the greatest average good? Or the greatest good for the "poorest"? Or should we adopt the non-sacrificial rule and accept bad situations for everyone just to avoid an even worse one for one single person?

2.3.3 Artificial consciousness

After treating AI as a moral agent, we shall discuss of AI as a moral subject: when a "Strong AI" is reached, when AI becomes aware, or at least capable of feeling pleasure and pain, raising a whole range of new ethical issues.

---

[4] *BOSTROM, Nick, 2014. Superintelligence: Paths, Dangers, Strategies. Oxford University Press.*

We first define Sentient AI (SAI) as an AI capable of feeling, perceiving and experiencing pleasure and pain. A further step is taken with Artificial Consciousness (ACn). Should we want to go beyond the intuitive understanding of consciousness that philosophers have suspected we share, despite their difficulty in putting it into words? We could follow the working psychological definition of consciousness (which, admittedly, is in itself subject to debate and shall be explored more in depth in our paper). In such a definition, ACn would be an artificial thinking entity possessing and integrating the faculties of *global workspace, information integration, internal self-model, higher-level representations,* and *attention mechanisms*[5]. Should we want to draw comparisons to the biological world, we define animals as possessing sentience and humans as possessing consciousness (although the limits may blur with the most intelligent species of animals, such as primates).

2.3.4 Other definitions

Throughout the document, we will be using the so-called "R. notation", inspired by Asimov short stories. In this paper, when naming a concept and preceding it with R. (short for Robot), we mean the equivalent of that concept for AI. For instance, "R. human rights" may mean "the equivalent of human rights for AI".

3. The Current State of AI

3.1 Historical development of AI[6,7]

One might keep in mind that AI's development is quite recent and has gone through cycles in people's expectation and technical progress. One can trace back AI to the early 50's, with the development of the first game AIs. In particular, one can mention Christopher Strachey's checkers program[8], finished in 1952, which was able to play checkers against a human using heuristics (as opposed to programs for games like tic-tac-toe where it is possible to enumerate all solutions explicitly).

In 1956 took place the Dartmouth Conference, which was the first conference about artificial intelligence, and during which the term "artificial intelligence" was created.

During the next 2 decades, there was a very quick progress of artificial intelligences, helped by a great optimism and lots of founding. The dominant approach at that time was to explore all possible situations in depth, using problem-specific heuristics to discard bad solutions. For

[5]REGGIA, James, 2013. T*he rise of machine consciousness: Studying consciousness with computational models.* Elsevier, Neural Networks.

[6] Wikipedia. History of artifcial intelligence.

[7] Interview with Prof. Boi Faltings. Head of the Artificial Intelligence Laboratory (LIA), EPFL.

[8]LINK. Programming ENTER: Christopher Strachey's Draughts Program

example for an AI playing chess, it would try to enumerate all possibilities for the next 3 moves, keeping the best one. To determine how good a move is, the programmer had to provide "heuristics" (i.e. an approximate evaluation of how good the situation is, for example if taking one of the opponent's chessman is positive, having the knights in a corner is negative, etc…). To avoid exploring too much possibilities, if the first move is already bad, the following ones are ignored.
Another important development during these two decades was natural language processing: being able to understand human (written) language (in the sense of extracting useful information out of it) and generating in return texts in human language. The chatbot ELIZA was finished in 1966, and enabled a written conversation realistic enough to fool some persons. However, it had no real understanding of the text: it mainly gave predefined answers (according to patterns and key-words) or transformed questions into affirmations based on grammatical rules. There was no notion of context or of the meaning of words.
Somehow, ELIZA demonstrated the superficiality of communication between humans and machines. With this chatbot emerged one of the first critics of AI. Indeed, already at that time, the author of ELIZA, Joseph Weizenbaum[9], wrote a book denouncing the dangerousness of his program because of the fact that people were not able anymore to distinguish computers and humans. People even thought that this program could become the first "electronic simulated brain"**.**
At the same period, were developed the first planification programs. For example, we can cite the first robot Shakey (1966 - 1972)[10] capable of reasoning about its own action. This marks the first great wave of excitement, with the emergence of the ancestor of expert systems. At that time, people and researchers hoped and thought that it would not last long till computers would be able to "do everything" and to exceed human capacities.

The second half of the 70's forms the first "winter of AI". At that time, most algorithms were based on enumerating all possibilities (even if some were discarded using heuristics). It was believed that the link between the complexity of a problem and the computational power needed to solve it was linear. But in parallel was discovered NP-completeness, which states that in fact, such approaches required a computational power growing exponentially with the complexity of the problem and there were simply no computers powerful enough to solve complex methods with this approach. Even if there was significant progress made in the 50's and 60's, it had been far slower than what had been announced and predicted, causing the founding for research to decrease drastically, therefore annihilating expectations, as well as the excitement connected to this research field. There was a big general disappointment.

At the beginning of the 80's, a new kind of AIs appeared: Expert Systems. The great novelty was that the "intelligence" was no longer embedded in the code itself, but in a database. On the one hand, there is an "inference engine", that can combine rules to calculate a result. On the other hand, there is a database of rules that are understandable by humans without

[9] Wikipedia. Computer Power and Human Reason - Weizenbaum.
[10] Wikipedia. Shakey the robot.

programming skills, and that could therefore be written by experts of a specific field rather than by programmers. For example, if one asked an expert system "is Socrates mortal", and there were the rules R1:"Man(Socrates)=true" and R2:"Man(x) => Mortal(x)", then the expert system could deduce that Mortal(Socrates)=true by applying successively R1 and R2. For the first time, we had a kind of "universal AI": the core program (the inference engine) can (theoretically) solve any problem, provided it has the right rules in its database. Expert systems were widely used in companies for specific tasks such as decision making. Nevertheless, if expert systems worked well on very domain-specific tasks, they couldn't be used for larger fields because it would have needed unreasonably big knowledge bases, which would have been too expensive to write and would have required more computational power than available. However, people still thought that this problem would quickly be overcomed and that logic programming would replace everything in order to produce a fully automated society.
The 80's also marked the apparition of multi-layers neural networks, which can be trained to learn a mapping between a multidimensional input space and simple outputs.

In the late 80's, started the idea of embodiment: abstract reasoning is only a very small part of human's intelligence and the most important is the reasoning and acting based on the perception of the world. Therefore, research began on putting intelligence into robots.
This led in the 90's to the notion of "intelligent agent", which is an autonomous "intelligence" that takes information from an environment and takes decisions based on them to optimize a goal. Note that the environment and the actuation might not be in real world, but can be digital, as for an AI doing automatic trading. Bayesian inference and stochastic gradient was also discovered at that time. This was the period were AI researchers made the biggest progress.

3.2 Current state of the art

There are today several approaches of artificial intelligence that are used in practice. We present the main ones in the following sections:

3.2.1 Optimization

One approach is to write the problem to solve as an optimization problem: the user has to write a "cost function" that evaluates how good a given solution to the problem is. There are then several existing algorithms (gradient descent, genetic algorithm, etc.) to try to find the best solution.
This approach is widely used in engineering (for example to find the optimal dimensions of a part) as well as in planning and managing (for example to create timetables and schedules).
There are however several drawbacks:

- The cost function is often hard to design, and can rarely be generated automatically.

- These algorithms usually find only local optimums (i.e. the solution they give is better than any nearly identical one, but there might be another one that is far better but totally different).
- They don't work very well when the number of parameters becomes too high.

3.2.2 Machine learning

Machine learning is a data driven approach. It tries to fit a mathematical model based on data which can then be used to make predictions for new data. There are three main kinds of use:

- Clustering: Identifying groups of "similar" data. It is widely used in online marketing to match similar profiles of customers and then be able for example to make relevant suggestions of other products the client might be interested in.
- Classification: Each sample belongs to a class (i.e. a "kind" of object). The program, during training, learns a model of the data. Then, when given a new sample, it can use the model to predict to which class it belongs. It can be used for example for a robot to recognize an object among a known set, or to perform optical character recognition (i.e. transforming a scanned document into text).
- Regression: this time, we no longer want to output just a category, but a number or a set of numbers, based on a dataset consisting in many situations with the corresponding output. It can be used for example to decide the speed of each motor of a robot according to the reading of its sensors, or to determine the best price of a product based on plenty of market parameters.

The drawbacks of theses methods are that they often require a huge amount of sample data (that is often expensive to gather)in order to work well and that there are often several parameters to fine tune manually.

3.2.3 Big data

Big data is essentially the same as classical machine learning, except that instead of having "few" (i.e. hundreds or thousands) of samples bearing lots of information each, we have far bigger datasets (millions or more samples) of samples embedding very few information. Therefore, big data focuses mainly in finding small trends in very big datasets that are cheap to collect (often coming from activity of people on the internet).

The main difficulties are that the data is often quite heterogeneous (i.e. the information available for each sample isn't the same), and that the computational power required is enormous (it often requires supercomputers and specific algorithms to make the program run in parallel on hundreds of processors).

3.2.4 Neural networks

Neural network algorithms try to imitate the neurons of the brain of animals (in a more or less precise way). Fed with large sets of data, a neural network can then be trained to learn the weights (the importance) of all the connections of the neurons it is made of. Once trained, one can then use the neural network to predict one or several outputs based on a given input.

The idea of artificial neural networks goes back to the 40's, but it is only in the last decade that we discovered how to train deep neural networks (i.e. with a big number of layers of neurons). We call it deep learning. Deep neural networks (DNN) are so efficient to analyze complex data like images that they replaced nearly all other image recognition methods. They are now also widely used in a lot of other applications.
However, they have three big drawbacks. First, they need lots of labeled samples. For image recognition for example, a DNN needs to be fed with hundreds of images of each object to recognize, and for each one a human has to indicate what kind of object it is. The second big drawback is that it is nearly impossible to understand what exactly a neural network learned. It is therefore impossible to predict what would happen if one presented an input which was not part of the datasets used for training. Of course this issue is not critical for many applications (like targeted advertisements put in place to manipulate public opinion) as incorrect outcome happens very seldom, but it is problematic if there are safety issues involved (like for autonomous driving systems, for instance). Finally, neural networks often require a lot of computational power, to the point that dedicated hardware for it starts to appear on the market.

3.3 Current research

Nowadays, AI is a very active research field, in terms of academic research as well as corporate research and development. Some of the most active and promising research fields are mentioned thereafter.

3.3.1 Optimization problems and loss functions

One aspect of AI which has been researched for a long time and which is currently well understood are optimization problems. Even if improvements are still possible, research is coming to a close. The focus is now set on how to formulate the goal to be achieved rather than to specify what the program should actually do. With this approach, the algorithm remains the same for completely different types of problems and only the loss functions (the criteria the optimal solution needs to satisfy) differ. Therefore, the real technical challenge is currently to formalize the criteria that must be optimized. However, this is not an interdisciplinary field of research and is still specific to problems.

3.3.2 Optimization and multi-agents

One way to optimize a problem more effectively or efficiently is to run several AI programs at the same time. The state of the problem is then no longer static for a particular AI program and an advanced collaboration between the AI programs is key in order to ensure a quick convergence of the optimization process. Research currently needs to make progress in decision making and in learning methods in the case of multi-agents system with a constantly evolving environment. One of the current research paths, as in many fields, is to draw inspiration from human behaviour.

3.3.3 Automatic tuning of parameters

One of the difficulties with many machine learning algorithms is that there are several parameters (called hyperparameters) that have to be set by the user and that strongly influence the quality of the results. For the moment, these parameters are chosen by learned guesses and/or by trial and error.

There begins to be some methods to determine the best hyperparameters, but they are not reliable enough yet to be used without human supervision. If we manage to find out how to tune these parameters optimally and automatically, it would make machine learning far more accessible, because one would no longer need to understand the algorithm to tune the parameter but could use it simply as a black box.

3.3.4 Selection of useful features

According to Prof. Aude Billard[11], one of the big drawbacks of current learning methods is that they remember everything, as opposed to humans (and probably most animals) that remember only the key information and discard the rest. The consequence is that most learning algorithms need to store and process huge amounts of data, making them very slow. Actually, many algorithms give good results when used on a powerful computer without time limit, but become completely unusable for real-time applications like the control of a robot, for instance.

More generally, any current type of AI is rather bad at extracting meaningful information. For example, suppose you have never seen a cat in your live. If someone shows you one cat that is sitting and tell you it is a cat, then you will probably be able to recognize also a cat that is walking, sleeping or that has a different color or is viewed from a different angle. Current AI would need to "see" at least tens, if not hundreds or thousands, of cats in all possible positions, with all possible colors, and from all possible view points to perform the same task and achieve the same results.

Therefore, the use of AI is for the moment mainly restricted to rather well known environments where it is possible to provide enough samples to be representative of all possible objects or situations. But as soon as it will be possible to extract useful features, the tech industry will have the tools to create much more "intelligent" AI programs and robots, capable of evolving in far more complex environments.

3.3.5 Understanding of learning and intelligence

One of the drawbacks of most AI methods is that it is very difficult or impossible to understand why they take a given decision, and even more difficult to predict what might be the output for a situation that hasn't been encountered by the algorithm before. Because we don't know why they fail on a given input, it is rather difficult to improve them and thus we can never blindly trust the results. This also creates the risk of ill working AI. Research has then been pushed to try to find ways either to prove that the outcome of the learning

[11] Interview with Prof. Aude Billard. Head of the Learning Algorithms and Systems Laboratory (LASA), EPFL.

algorithm doesn't break some predefined "safety" rules, to visualize what was learned, or to make the algorithm "explain" how he took some decision.

According to Prof. Billard, there is much hope of progress in the understanding of AI and, as a good side effect, also of human and animal intelligence. However, other researchers like Prof. Faltings think that the hype and the expectations around those AI are again too high. Indeed, they think that deep learning, for example, won't be able to solve every problem even once greatly improved thanks to the amount of data and the computational capacity that we have nowadays. In a vast majority of cases like shape and image recognition, deep learning has good results, sometimes even better than what humans could do. But it does not work well in many other applications like speaking language and it is always possible that they fail completely in an unexpected situation. Moreover, even though deep learning algorithms can often "solve" difficult tasks, the error rate is about 20-30%. This explains the prudence of these researchers with respect to deep learning as such results do not mean anything since we could achieve the same outcome by following random phenomena.

3.3.6 Data acquisition

With the raising concerns about privacy, the big amount of data gathered for learning techniques needs to be handled with great care. Moreover, since learning algorithms need so much data to be effective, they must come from a large number and variety of sources. Therefore it is very important to be able to gather data of good quality and accuracy, as well as to be able to verify it. One way would be to use techniques from game theory that would improve the quality of data and ensure its correctness, for example by rewarding sources providing correct data.

This reveals the (potentially ethical) challenge of defining the criteria used in the decision-making process which in addition must be reliable and verifiable, especially if in the future such decisions will be made automatically by a software.

3.3.7 Trends for the next years

To give an idea of what to expect in the next years and decades, we will present a few figures from *When Will AI Exceed Human Performance? Evidence from AI Experts*[12] by K. Grace and al. published in May 2017. This study is based on the survey of 352 searchers in AI that were asked about when they think that AI and/or robots can perform different tasks.

Half of the interviewed researchers believe that:

- In six years, robots will be able to fold laundry as well as and as fast as professional humans.
- In eight years, AI will be as good in translation as amateur humans and will be able to provide phone banking service as well as a human.
- In nine years, AI will be able to recognize the type of an object based on an unique sample and to read a text aloud with the same skills as an actor.

[12] GRACE and al., 2017. When Will AI Exceed Human Performance? Evidence from AI Experts. ArXiv.

- In ten years, AI will be able to write a high school essay as well as good students (and pass the plagiarism tests successfully) or explain its choices in games (like chess) in a way that humans can understand.
- In eleven years, AI will be able to produce a song imitating a given artist in such a way that nobody can recognize that it isn't the artist who did it.
- In twelve years, a humanoid robot will be able to win a five kilometers race against a human in a city.
- In 33 years, AI will be able to write a best-seller.
- In 36 years, robots will be able to work as surgeons (and outperform humans).
- In 43 years, AI will be able to do math research (i.e. find new theorems and prove them).
- In 45 years, AI will outperform humans on all tasks.
- In 120 years, all (current) human jobs will be automated.

These numbers must however be interpreted with much care as there isn't a consensus between researchers. Some researchers, such as Prof. Faltings, think that for example there won't be completely autonomous cars before five years but that it will certainly be achieved in the next fifty years. Another example is Prof. Guerraoui, Head of the Distributed Computing Laboratory at EPFL, who thinks that the proofs of theorems will not be within the reach of AIs (or at least not before a very long time). Research and development of AIs remain at the moment very domain specific and it could stay like this for a while because researchers still have an enormous amount of work to provide for the generalization of AI. However, there are some attempts to address quite general problems. For example, teams of thousands of researchers at Google are trying to connect speech and vision. Unfortunately, due to a lack of manpower, this kind of work is out of the reach of universities for now.

Based on the previous discussion, it is possible that neural networks (and those kind the current forms of AI mentioned above) will not be part of the future of AI. This presupposes that research will soon discover another method to achieve the generalization of AIs. It is indeed very important to know with precision how learning algorithms learn and to be sure that their outcome is correct, as this is not the case for all applications at the moment.

4. Superintelligence: AI as a tool

Despite the discussions on AI development mentioned above, it is a fact that Artificial intelligence algorithms are becoming more and more powerful and outperforming humans for many specific applications. They are not only becoming more "intelligent", but they are also becoming able to cover larger and larger fields of competences. If we suppose that this trend will go on, in a near or distant future, there will be AIs able to outperform mankind in a whole field of expertise. It will be precisely from that moment that they will become SIs.

4.1 Genesis of the SI

Nowadays, there are many specific applications where artificial intelligence modules are used to improve performances of high-tech devices. The next step that we can consider will probably be to connect some AIs together in order to create clusters of AIs able to cover even larger fields of knowledge and expertise. From that point, we are not very far to connect enough of them (or to improve them) in order for them to be able to solve nearly any task better than humans. At this stage, we can start to call them super intelligence (SI): not only is it able to perform specific tasks better than humans, but it is also able to make connections between totally different fields, that a human would never had thought of, and that might turn out to be highly beneficial.

In the rest of the report, we will mainly talk about the SI in the singular form. However, this is not (at least at the beginning) a unique SI that would act at the level of humanity or global society.[13] There will probably be several SIs that would collaborate or challenge each other. Nevertheless, the aspects concerning these SIs will be similar and, for this reason, we will use the singular form.

The genesis and the development of the SI will probably make use of new technologies and paradigms such as IoT (Internet of Things), ubiquitous computing, cloud computing[14], data mining approach, etc. Moreover, global and borderless databases (for well-structured data) and knowledge bases (for any other data) will probably be created (some already exist) in order to give the possibility for SI to learn from them, improve itself and become more and more effective and efficient. It is precisely the mechanism of deep learning introduced above. Learning algorithms initially proposed by humans will then be improved by the SI itself. For example, there are already genetic algorithms and artificial neural networks that are developing in this direction and that will allow the SI to evolve through mechanisms of "artificial selection" (which is similar but more effective and faster than natural selection). Once such a process has begun, the speed of artificial selection will gradually increase. This would lead to an explosive intelligence that would allow the original AI to evolve into an SI that would perfect itself (as anticipated by the philosopher Bostrom).[15]

However, there are some algorithms that can be implemented in order for the SI not to avoid human intervention or not to create situations where it can't be stopped by humans. This is called "the safe interruptibility".[16]

One important point has to be considered while conceiving the SI: the impact it will have both on individuals and on society in general. There are many situations that should from the start be, if possible, reduced or avoided by including specific features (protections) into the SI. For example:

---

[13] This has been validated by Prof. Billard and Prof. Faltings during the interviews we conducted (cf. Annex B)

[14] 2017. Artificial Intelligence for Cloud-based Internet of Things (IoT). Call for Papers Elsevier.

[15] BOSTROM, 2009. Superintelligence. Oxford University Press.

[16] EL MHAMDI, GUERRAOUI, MAURER, 2017. Dynamic Safe Interruptibility for Decentralized Multi-Agent Reinforcement Learning. Conference on Neural Information Processing System.

- A protection against using the SI for bad purposes (within the limits of what is feasible during conception and implementation). It will be essential to be able to anticipate these risks as early as possible and to properly define what might be "bad".
- A way to make sure the SI doesn't discriminate some categories of groups or individuals, whether on base of racial, religious, social, wealth or any other unfair criterion. Such attempts already exist and are in full swing in the social networks like Twitter or Facebook.
- The ability to realize when a human is trying to manipulate it. In this sense, we have to be very careful because this can make the SI conscious of the fact that it can oppose and disobey humans.
- A reliable way to auto-check that the result isn't wrong or misleading.

4.2 SI: stakeholders and their implications

Through its action, SI itself might be considered as the main actor. However, there are many human actors that are also strongly involved or concerned:

- The creators of the SI. They spend plenty of time to create the SI, and might (or not) get a share of its success. Many might consider them as guilty of any misdeed committed by or with the SI. However, once launched, they might well lose any control or influence about the evolution of the SI. These are the "first" creators of the SI because, from a certain level of development, the SI will maybe be able to reproduce itself.
- The persons propagating the SI. They sell, lend or share (with or without conditions), promote, maintain and update the SI.
- The end-users (the persons using the SI in order to get solution to problems). They do not necessarily understand how the SI works internally but are able to make use of it.
- The regulators and legislators. They are in charge of imposing and enforcing legal measures and norms in order to avoid dangerous SI or dangerous uses of SI. Unfortunately, experience shows that sometimes, or often, the adaptation of the legal framework is rather slow and it lags behind the rapid progress of new technologies.
- The supervisors. They are in charge of checking that the output of the SI is, as far as possible, always acceptable and make sure no suspect request is processed.
- The people who undergo the (direct or indirect) effects of SI's action and decision. If SI is used to find the solution to a problem, this solution will probably be applied or used. However, it will have different impacts on different people.

For example, in a medical intervention managed by or involving the SI, human actors could be the following. The creators could be a mixed team of medical professionals and multidisciplinary engineers. The end-users could be the medical teams. The regulators could be the SI experts group of the Ethics Committee of the Ministry of Health. The supervisors could be the hospital's medical ethics team and finally those who undergo the effects of the SI's action could be the patients.

4.3 Social impacts of SI: 3 different perspectives

4.3.1 *SI: a fantastic tool for the greater good*

SI is very powerful tool, able to solve any problem better than human would have done. Therefore, if we apply it to answer questions related to the well-being of people, we can get much better solutions than if a human had tried to solve the same problem. There is an infinity of applications with big impact on the greater good. We could for example apply it to optimize agriculture in an environment-friendly way, to find a vaccine against AIDS, to find out the best tax system or simply to optimize public transports. And if the computational power of the SI isn't enough, we could just ask it to develop a better version of itself. Whatever problem humanity might be facing, SI will solve it at least as well as the best specialist: the result for the greater good can only be positive.

4.3.2 *SI: a mere tool, neither good or bad in itself*

But what if someone (or some entity) ask some selfish request, for example "how to make as much money as possible"? The SI will of course answer as precisely as possible, without caring about the consequences. Similarly, the SI could be used to prepare and wage wars with surely disastrous repercussions. In fact, among the main investors in the development of the SI are precisely the military. Even worst, if some terrorists asked how best to destroy the world, they would get a good answer to their request! In the same line of disaster scenarios, totalitarian governments may be tempted to rely on the SI to manipulate the masses and to monitor, control and discredit the opponents.

We have seen that SI can do a lot of good, but also a lot of evil. As a matter of fact, SI is merely a tool, even if it is probably more powerful than any before. Like any tool, it is neither good nor evil in itself, only the use that is made of it counts. If someone uses it for the greater good, then it can yield fantastic results, but if used with bad intentions, it might lead to horrors beyond imagination. This is the perfect example of the dual use dilemma.

4.3.3 *SI: the end of human work - for the best or the worst*

Another big question is the social impact of SI on our societies. As a matter of fact, a SI would be more "intelligent" than humans and therefore more fit to solve any intellectual task. Moreover, they will probably only need some pieces of hardware to run on, and will therefore be far less expensive than humans for realizing intellectually complex tasks. As for manual tasks, robots are already way faster and more precise than humans, but still lack the "intelligence" to adapt dynamically to new tasks, which restrains their use to repetitive tasks for which it is worth to pay someone to deploy and program them for the specific task. But by embedding an SI, these drawbacks would disappear. SI seems to be able to perform any task (if embodied in the right robot) better than humans. So why would anyone want to employ a human when a SI can do better for cheaper (on the long term)? The emergence of SI might

well mark the end of human employment. Is it a good thing? It depends of how things are managed.

On the one hand, it would mark the end of human labor, that is generally considered as something we only do for the payment but that is a pain. In fact, the French word for work, *travail*, comes from the Latin *tripalium*, an instrument of torture. Therefore, if well employed, SI could deliver mankind from any labor and enable him to have complete freedom about what he wants to do of its time. It might be the beginning of a society of pleasure and leisure. All the goods and services needed (or wanted) would be provided by SI, so there would be no use to work. However, those wanting to work could of course do so, on whatever subject they want.

On the other hand, the ideal situation above yields only under one condition: that the SI are used only for the greater good and that their production benefits to all. But what if it is the exact opposite, and persons or companies caring only about their own interest control the SI? They would probably replace the human workers by SI that are more efficient. But they would probably not share the benefits with those they just sacked. Therefore, a very small minority would own the SI and all the production they could secure, whereas the big majority would not only get no share of the benefit but would also be in total incapacity of finding any job (because they wouldn't exist anymore). This would then increase inequalities to a level unseen before: a small minority with nearly infinite wealth and the rest unable to earn their own livings.

As we have seen, SI could make wonders if we find a way of making sure the benefits are equally shared, but could also lead mankind back to its darkest periods in terms of inequalities. The challenge is therefore to find some way to enforce this fair sharing, which would probably not occur spontaneously in our capitalistic society.

4.4 Limitations and obstacles: potentials errors

According to Prof. Faltings, some fundamental problems will always remain and will not be able to be solved by SI. Those problems are, for example, induced by human nature (jealousies, conflicts, comparison with others) or related to the competition for resources. For a while, the SI will remain a tool. It will help us do tasks but it will not solve the fundamental problems of humanity that have remained the same for thousands of years.

Moreover, how extraordinary SI's capacities might be, there is always a risk of error or misbehaviour. Even if during the conception and deployment of the SI features were designed and implemented in order to insure fair and reliable action, some unpredicted situations might occur in practice. They might be aspects that were simply forgotten to take into account. They might also be side effects either due to hazard, to a complex combination of facts beyond human (and SI) comprehension, or to some malicious actions. The case of the share of moral

responsibility and liability if such incident happened, should be defined as precisely as possible, even before the SI is put into service.

In order to exist, the SI may use huge data and knowledge bases. Therefore, it could exploit this information in a transparent manner for more or less honest purposes and that could affect the privacy of individuals or groups of people. For example, a person whose "private" data is used by the SI should at least have the right to know what specific information about him/her is stored, for how long (respecting the right to be forgotten), who has access to the data and for what purpose. Moreover, transparency is useful because, on the one hand, it comes from the SI's learning and development process and, on the other hand, it would ensure the security of the SI against the malicious actions of humans trying to corrupt its integrity.

Given the aspects discussed above, it can be seen that the use of the SI in its "raw state" is subject to limitations, or may even be dangerous. Therefore, additional elements should be added to it in order to make the SI's behaviour more moral.

5. Artificial Conscience: AI as a moral agent

We developed in the previous chapter the challenges of SI. Let's now continue the discussion by considering an improved version of the SI: Artificial Conscience (AC). AC is essentially SI with an additional characteristic: morality. The point of this additional step in the development of an SI is to be sure that they can be trusted to take undisputable decisions.

AI could take into account moral judgment through adequate constraints. AI is essential in such an approach because conventional computer programs can not take into account such fairness integration. We already have the necessary technology to implement it but the main challenge to be overcome is the efficient and correct formalization of these constraints into logic. Moreover, we must pay attention to the criteria chosen for optimization in order not to discriminate certain stakeholders, for example. According to Prof. Faltings, implementing a morality in an SI normally goes against the principle of optimization (needed in such SIs) because such an implementation introduces supplementary constraints and would then affect the efficiency of the algorithm. A particularly appropriate example is the one of autonomous driving systems that must make critical decisions in situations where casualties are inevitable in a very short amount of time (a few tens of milliseconds, depending on the speed of the vehicle). We will further develop this particular case in section 5.2.

As another motivation to the following discussion, let's see what could go wrong in a learning algorithm if no moral rules were defined. A company wants to reduce its HR costs by using an algorithm to do a first selection of the applications before the interview. The company

feeds the algorithm with resumes of their current exemplary employees (the single resumes they have free access over) to train it to recognize a potential good candidate. They currently only have Caucasian employees but are now required to diversify their manpower. After a few weeks, they are disappointed to notice that even if they were accounting for 50% of the applicants, only Caucasian applicants had been selected by the algorithm for an interview. Indeed, if you train your algorithm on a random corpus without any additional rule, very high are the chances that the omnipresent *clichés* will be learnt by the algorithm. A very undesirable side-effect which is just reinforcing existing biases. The point of the moral rules is to be sure that the algorithm won't act as a common human, but as a perfectly ethical one. On that particular example at least, AC would do a better job than a human to pick a candidate for a job interview based on his/her resume. This is because the subjective point of view of the human is here a handicap as biases are unconsciously affecting our judgement.

5.1 Teaching morality to SI

There are currently two main approaches to instill morality in SI. The first one is to precisely define the behaviour to adopt for each possible situation the SI might encounter; the second one is to only define general paradigms and let the SI take the decisions by itself, based on these basic initial rules.

Being sure that the SI will behave as we naturally and implicitly intend to as human beings is the ultimate goal in teaching morality to SI. Unfortunately, it is illusory to think that we can foresee all possible situations SI might encounter and come up with a clear and definitive outcome for each of them. First of all, there is no consensus among human beings on what is right or wrong: believes might be different based on gender, cultures, social class, etc. This is also very difficult to program human interactions and morals, notions that are for instance often subject to change over time.

Computers are binary while humans are moderate; the translation between both worlds is not trivial. If it was easy (or even possible) to define in a deterministic way how people should behave, the world would be very different: no wars, no racism, no homophobia, etc. Conflicts would be totally eradicated.

The partial solution, and the approach currently used, is to find higher-level rules to guide the SI in its daily decision making. SI will still require them to be defined very formally and not contradict with each other, but this non-trivial task is still more feasible than the explicit method. In addition, this method seems to be more logical and intuitive. Indeed, children learn how to grow and interact socially by mimicking their parents and other relatives' behaviours. The programmers of the SI would then take the role of the parent: teach a set of basic rules, be sure that the SI is indeed following them as intended and potentially clarify them as particular and error-prone situations occur.

But can we really give this power and responsibility to a single individual or even a group of random people, depending on what will be the purpose of the SI? Here, the "no" prevails for multiple reasons.

First, there might be conflicts of interests. The masterminds of these moral rules might be tempted to design them such that the resulting AC will act in their own personal interest. Then, they might not be knowledgeable in ethics and might not take appropriate decisions. But most importantly, these moral rules should be universal and unique among the world.
We still have to answer who will be responsible to design the moral rules of AC, as it has to be thoroughly thought. One possibility would be to poll the entire world population on different case studies to grasp the general opinion of the human beings and use it as a pseudo-consensus. Unfortunately, this would not necessarily give an ethical outcome as not all humans are sensitive to ethics. For instance, if they had to choose between killing a prisoner and a top executive, they would most likely choose the prisoner, even if this is unethical to consider that human life can be discriminated. Another possibility would be to equally represent each community of human beings by a single expert in ethics. But of course, this is very difficult to identify such communities. And even if such a commission could be set up, another problem would be to have the authority and the power to enforce the application of their decisions. We will have to use our experience with the United Nations which currently struggles in their governance to avoid repeating the same mistakes.

5.2 The example of the autonomous car

One of the hottest topic on morality of AI is in self-driving cars. Even if we can decrease a lot the number of accidents with self-driving cars, fatal situations will still arise. Currently, with conventional cars, most of the time the driver doesn't have time to take a thoughtful decision and might just try its best to reduce its own injuries, at best. Self-driving cars, however, can't stay in uncertainty and will need to do a choice. This choice will first be based on the context of the accident and the basic rules of morality received at its creation, hence the need to thoroughly think about the strategy we have to adopt in these extreme situations.
The German government has been the first to create an ethics commission for automated and connected driving. The report of the commission[17] contains, among other things, twenty rules that must be followed while implementing self-driving cars. The two most important state that human life is more important than animal or property, but that all human lives are equal.
For instance, this would not be acceptable to choose to kill two old women instead of a pregnant woman pushing a stroller. Under these circumstances, the car has to determine which scenario minimizes the overall harm.
But who would buy a car which can potentially kill him/her? We have here a very good illustration of why moral rules cannot be taught to AC by a random set of people. If only people who have interest in the car company are implementing such rules, this is clear that the personal individual good of the occupant of the car will be prioritized over any other outcome like the death of multiple children in a playground due to a break failure.

[17] GERMANY, 2017. Ethics commission - Automated and connected driving. Federal Ministry of Transport and Digital Infrastructure.

5.3 The humans in an AC society

In the chapter 3, we wondered if developing SI and integrating them in our society would bring enough benefits to counterbalance the potential drawbacks. This question is not easy to answer without a deep cost-benefits analysis. But we can at least partially circumvent it by considering AC instead of SI. Indeed, providing SI with morality can solve a lot of problems of its potential adoption in our society.

If built properly, an AC would be at last equivalent to a human in terms of labor and execution of tasks. It would then make absolutely no sense to stop the contribution of these super-humans. If AC are not currently invading our society, it is not only because of our lack of success in building such agents. Regulations refrain a lot the progresses in AI. Once the regulators will be confident that AC are totally capable of operating in their assigned fields without compromising the overall good for society, they will proliferate at large scale and we will have to redefine new roles and goals for humans in our society.

Robots can already automate a lot of mechanical tasks and already take some jobs to the humans. The production factories and the checkout line of groceries stores are good examples. This is why we start to speak about unconditional basic income. With AC, way more jobs will be closed to humans and the employment rate will go through the roof. We will then have to seriously consider the unconditional basic income, or only people working in very specific sectors (which remain to be determined) will be able to sustain a decent quality of life.

These sectors will most likely include IT engineers building and maintaining AC. But based on our discussion, we can also assume that people with very good ethics will be very valuable to be able to keep a good collaboration between human beings and AC.

6. Artificial Consciousness: AI as a moral patient

We possess a common acceptance of the fact that consciousness gives a basic dignity from which inalienable rights naturally follow - as first hinted at by certain Italian Renaissance humanist authors[18] [19], then theorized (and attempted to be implemented) by Illuminism [20] [21] and finally codified by the UN Universal Declaration of Human Rights in 1948.

In the same way, the western collective is growing a common acceptance that animal sentience gives them some Animal Rights, as natural and intrinsic as Human Rights, if subordinate to them and ignorable for the sake of human development and comfort.

[18] Marsilio Ficino, *Theologia Platonica, "Man is similar to God" passage*, 1482
[19] Giovanni Pico della Mirandola, *Oratio de hominis dignitate*, 1486
[20] John Locke, *Two Treatises of Government*, 1689
[21] French National Constituent Assembly, *Declaration of the Rights of Man and of the Citizen*, 1789

Logic would demand that, should AI gain sentience (although this is, admittedly, such an alien concept that the following discussion will be by nature quite theoretical), they should possess rights similar to those we attribute to animals: "R. Animal Rights"; that, should AI gain consciousness, they should be recognized rights comparable to those we attribute to humans: "R. Human Rights"; and we should not exclude the possibility that in a certain future, AI could be developed to a further step of awareness and sensibility, to a more advanced or evolved state of consciousness (similarly, we should think that man could also reach this state through transhumanism) - and as scary as the thought may be, logic imposes that would grant them rights beyond that of human.

The immediate ethical questions and issues that arise from this train of thought are obvious, especially in the light of the historical knowledge of how slow we have been to recognize rights to ourselves (the occidental man, late 18th century[22]), never mind to those a bit different to us (the whole of mankind, half of 20th century[23]). We shall see how much time it will take to recognize machine rights - if such a notion even makes sense, of course.

Beyond this obvious - but not any less complex - ethical issue, many more can be developed. This chapter deals specifically with ACn ethical issues, but many of the reflections and observations can similarly be applied to SAI.

6.1 ACn ethical issues

First of all, do we even have a right to create consciousness, as a god would? How should we create it, with which features (e.g should it even be able to feel pain?), drives, ambitions? If it is indeed possible to create advanced AI without consciousness, what added value does it give to justify its addition?

Furthering the discussion on R.Human Rights, how do the features and character of a particular ACn influence the rights it possesses (e.g. if a particular ACn was not given a survival instinct, does it make sense to recognize a right to life for it?)? Can the private possession of ACn be morally justified? How about the tampering (e.g. erasing/altering the memory, rewriting the personality, forcing certain emotions), damage or destruction of ACn?

On the other side of the coin, what about the responsibilities of ACn? Can a consciousness fully architected and created by another entity bear responsibility or blame? How do we distribute blame or responsibility (to the developers, to the educators, to the people and ACn it came in contact with) in the case of faulty or rogue ACn? Does it make any sense (moral or practical) to establish a crime and punishment system similar to that of humans?

[22] *Declaration of the Rights of Man and of the Citizen*, 1789
[23] United Nations General Assembly, *Universal Declaration of Human Rights*, 1948

What about the social impact of ACn, when AI becomes a valid and fulfilling - and even customizable! - companion for any human wealthy enough to buy one? Is there a risk to break down human links any further than they have been broken by modern society?

And finally, due to how our own mind and morality works, should any of those rules be different for ACn in human-like bodies - for androids?

6.2 Creation of ACn

Starting from the top, the first topic to be discussed is the ethical concerns relating to the creation of ACn. Leaving to the side the more theoretical (perhaps even theological) discussion on whether we even have the right to artificially create conscious beings, we will concern ourselves with the practical side of whether we should or not create ACn and, if we should, on what terms and with which features, traits and characteristics.

Contrary to SI and AC, we see no great, revolutionary changes or improvements the addition of a consciousness to AI would bring upon mankind. We would argue that, should ACn be developed, it would be for one or more of the following reasons:

1. For the pure sake of scientific invention and development;
2. Because the features which make consciousness (§2.3: *global workspace, information integration, internal self-model, higher-level representations,* and *attention mechanisms*) are otherwise useful, and their implementation inevitably or accidentally creates consciousness;
3. For use as companion AI (or similar tasks requiring a "human element, e.g. assistance to the elders, medical care and rehabilitation, etc.), as interaction with a conscious being is intrinsically perceived as more meaningful and satisfying than interaction with a simulated consciousness.

I will also mention the hypothesis that creating happy ACn might have intrinsic moral merit: that ethics would dictate that we *should proactively create as many as happy as possible ACn, for no other reason than the creation of happiness*. We did not include this train of thought in the list because, even if it may have merit from the point of view of a logic detached from instinct and reality, it will surely not be followed - as its acceptance would depreciate and make the pursuit of human happiness almost obsolete (if ACn happiness is worth as much as human happiness and is easier to achieve and even "mass produce", logically we should devote our efforts to achieve it), nullifying our raison d'etre.

Whatever the reason for creating ACn may be, the decision on whether or not to create it and how to create it will be in the hands of those at the forefront of AI development: governments

(military R&D sections, in particular), leading enterprises, and those that might influence these two - such as lobbyists or private investors.
After ACn becomes a reality (and after we realize it has! Telling consciousness from the sum of the elements that lead to it might prove tricky); or rather after its creation or the possibility of its creation enters the public consciousness, legislation regulating it will have to created by the legislators, under the inevitable influence of the voters and the lobbies.

Such a legislation, regulating the features of ACn (e.g. should we ban the implementation of a desire to pursue freedom?), is necessary on one hand to properly bring about the benefits of ACn (which might be the only form of advanced AI!) to society, and to safeguard the happiness, or at least the wellbeing, of ACn.

Failure or delays in implementing regulations would not just jeopardize proper distribution of ACn benefits and ACn wellbeing.
Indeed, as tech and security experts warn us today[24] [25], computer systems pose an enormous threat to personal privacy and, as a consequence, freedom - in no small part due to how legislation is murky, and how the control entities are neutered and ineffective. One could only imagine how these threats could translate to a world with ACn - which are implied to be in frequent contact with us, and with which there could be potential of developing an emotional link.

6.3 Responsibilities towards ACn

Regulating ACn creation and characteristics would of course only be the first step in order to guarantee a fair and good use of ACn. The next one would be regulating the treatment of ACn: legislation and norms dealing in particular with the almost inevitable relation between the owning of ACn (ACn as property: they are produced, thus are going to be property - or there would be no economic interest in developing and commercial interest in producing them) and the fair treatment of ACn (ACn as a conscious being).

Establishing rules and limits on modifying, tampering, damaging, distressing, disposing and destroying ACn will no doubt be a legislative nightmare, especially considering the novelty of the matter at hand and the fact that different models of ACn, possessing different feelings, capacities, drives, and motivations have no reason to be subject to the same rules.

[24] Douglas Crawford, The Intel Management Engine - A Privacy Nightmare, 2017. One can only be surprised (or alarmed) on how such a glaring and ubiquitous breach of privacy and security receives such little coverage and attention.
[25] John Naughton for *The Guardian*, *Google, not GCHQ, is the truly chilling spy network*, 2017. This is one example among many: leaks imply that virtually all the social and messaging system either collaborate with security agencies or are compromised by them.

Indeed, one could argue that jurisprudence bases itself on the will to not violate the basic, natural instincts and drives of man (which motivate his rights)[26]; and that, subordinately to this, modern right also aims to respect and allow for his ambitions to be pursued in a fair context.
To exemplify this concept, the prohibition to kill has always been part of any legal system, due to the powerful instinct of survival and will not to be killed; social convivence could not be possible without the security brought on by the guarantee from the state and community that no one is going to kill you. As for the second part of the statement, most social movements of the last two centuries have had for aim to grant more equal treatment and opportunities for various mistreated social groups.

To carry over this concept to R. rights, one could argue that the rights granted to them should correspond to their R. instincts, to what causes them distress, to their drives. If an ACn was not developed to have a desire to pursue freedom, it would make no sense to grant it the right to be free; if an ACn possesses a strong survival instinct, and threats to its life put it in distress, it should have a right to life, but if an order from its owner overrides that instinct and nullifies that distress, it could be allowed for the owner to override the ACn's right to life.
Necessarily, not all ACns will have the same drives and instincts; as such, the same rules and rights should not apply equally to them.

In addition to being a legislation hard to elaborate, it might also prove hard to implement and enforce. How can the law guarantee the good treatment of ACn placed in private buildings? Not just cases of abuse in a domestic context (and one needs only look at human domestic violence statistics to see how hard it is to notice and track such cases); but, for instance, any industrial ACn could be subject to stressful misuse.
The possibilities of abuse are too many to list; and we do not yet see a way to reliably and systematically discover and report them that doesn't involve a blatant violation of privacy or the facilitation of such[27].

While a daunting tasks for legislators, and a high responsibility for the voters and lobbyists that influence them, it is key in ensuring fair ACn treatment; and maybe, as a consequence, good ACn behaviour. Furthermore, the same legislation could also guarantee good development and production standards to ACn in order to prevent defects; while that would undoubtedly raise costs (and as a consequence further restrict the number of people with access to this technology), the end product would be more reliable and of higher quality.

[26] John Locke, *Two Treatises of Government*, Ch. II, sec. 6,: "[...] no one ought to harm another in his life, health, liberty, or possessions." Such would be the natural rights of man, according to Locke, anterior even to any man-made legislation.
[27] And indeed this could serve as a pretext to justify a continuation of the contemporary privacy violation described in §6.2.

6.4 Responsibilities of ACn

Another of the pillars behind (modern) jurisprudence is that its purpose is to deter crime through the threat of punishment[28]. Could we envisage a similar system for ACn, to encourage and enforce good behaviour? After all, with rights come responsibilities; with R. rights could come R. responsibilities, and R. crime and R. punishment.

Or maybe it would make no sense to punish, or even have a notion of "responsibility" and "blame" for a consciousness fully architected and created by another; perhaps, in case of bad behaviour of ACn, the only blame is to be put with the developers and producers.
Such thoughts are not new to us; notably there has been many a theological discussion on the topic of the responsibility and sin of a man created by God[29].

Most of these discussions ended with the conclusion that man still bears responsibility for his actions and blame for his crimes; and those that didn't were all but ignored - after all, to void the penal system (in many historical cases claiming to be the direct extension of the will of God) would mean to destroy society.
But would all of this be necessary with ACn, if we make the hypothesis that careful "programming" could make them simply *unable* to commit crimes?

Perhaps an R. criminal system (or an equivalent moral system[30], or both) would only be useful as a redundant safeguard. On the other hand, its implementation - either by legislators or by the fabricants themselves - could help with the integration of ACn in human society - being subject to similar rules, to a similar system, could help users relate with and accept ACn more easily.

6.5 The case of androids

Many recent studies in neurology have focused on the role of mirror neurons and shared neural activations (following Hebb's rule: "Neurons that fire together wire together"[31]). Some of those studies highlight the role of those in the empathy process; and, perhaps, the subconscious influence they have on the moral system and values[32] .

---

[28] Cesare Beccaria, *On Crimes and Punishments*, esp. Ch. XII and XXVIII, 1765. Considered by many to be at the base of modern law.
[29] See for instance the discussion around the Fall of Man and the Original Sin.
[30] Notice that in this context the difference between criminal and moral system would be that in the first the punishment is external, delivered by a third party; while in the latter, the "punishment" would be internal and self-served, expressing as negative feelings of "regret".
[31] Cit. Donald Hebb, 1949
[32] C. Lamm and J. Majdandzic, The role of shared neural activations, mirror neurons, and morality in empathy--a critical comment, 2015

On a more sociological note, there has been over the last century a constant debate on whether violent representations (such as those found in movies[33] and video games[34]) makes people more aggressive or even prone to violence; an idea strikingly similar to the enduring and ancient belief that public displays of indecency corrupt public morals; and one can always remember how disfiguring or desecrating a representation of something sacred (religiously or morally so) is and has been seen at best as in bad taste, and at worst as "blasphemous" and illegal[35].

These beliefs are indeed quite similar: that witnessing a taboo action or even witnessing or acting out a representation of that action subconsciously weakens the taboo, makes you more likely to commit that action, and should as a consequence be avoided.

The advent of ACn, and in particular of androids - humanoid robots - will carry with it the possibility of legally harming something very, very closely resembling humans in both appearance and -for the first time- behaviour. Under the light of these previous examples: should we consider stricter restrictions on the treatment of androids than we would impose for non-humanoid ACn? For instance, should "simulating" torture on an android that doesn't feel pain or distress from it be forbidden?

It's up to legislators and public opinion to decide; certainly studies from behavioural and neuroscientists will be needed. For once, the potential victims of the lack of good legislation on this topic wouldn't be the ACns, at least not directly; but rather human society at large, through a potential weakening of public morals and collective taboos.

7. A Preliminary Ethical Evaluation in a Scenario-Based Approach

For the moment, we mainly assessed the different kinds of AI separately, considering them as established. However, none of these AI will occur instantly: there will be an evolution toward it, and it is during this evolution that will be determined how these inventions will be used and who will benefit from it. What's more, there are several possible scenarios, between which it is not yet possible to know which one will happen.
We will therefore adopt a scenario-oriented approach to explore several possible evolutions.

[33] Keith Perry for the Telegraph, *Watching violent films does make people more aggressive, study shows*, 2014
[34] Alexandra Sifferlin for The Time, *Violent Video Games are Linked to Aggression, Study Says*, 2015
[35] Even in recent years, and western countries: see for instance the prosecution and 12 month sentencing of Simon Glerum (33) in the UK in 2017 for importing via mail a "childlike sex doll". Children being sacred in our society and pedophilia being one of the strongest taboos, even "victimless" expressions of it are persecuted. Lizzie Dearden for The Independent, *Man sentenced for importing childlike sex doll from Hong Kong*, 2017

7.1 Scenario I - SI in the hands of big companies

For the moment, most of the powerful AIs are developed by big companies like Google or Facebook. There is therefore a rather high probability that such companies will be the first to develop SI.

By definition, the first SI will be better than humans in (nearly) all intellectual tasks, and in particular in developing new and more powerful SI. The first company to have an SI will therefore have an head start becoming bigger and bigger as time passes.

We can therefore assume that one big company will have a monopole in the development of top SIs. This company then has 2 interesting options to take profit of this SI.

7.1.1 SI accessible to big companies

The first use is to sell licenses for this SI to others companies. These companies will probably use the SI in order to replace humans for intellectual tasks because it is cheaper and yields far better results. So, on the one hand, these companies will be able to sell far better products at lower costs, which will probably cause the other companies to close because there are no longer competitive. On the other hand, most of the employees doing intellectual tasks in those companies will probably be sacked because there aren't useful anymore. We can therefore expect most intellectual jobs to be destroyed within a few decades (if not less) either because the smaller companies close, or because the task is achieved better by SI.

What's more, all the manual tasks are in even greater danger: during the last decades, most purely manual tasks were replaced by robots that are cheaper. The tasks that remain manual are mainly those which are performed in to low quantity or that are too diversified to make it interesting to program a robot to do it. So what saved manual jobs so far is that there still need a bit of brain and adaptivity. But once we have a SI (or maybe even before, with simply a better AI), it will be able without difficulty to control the production robots in a smart enough way to perform complex tasks that aren't repetitive. We can therefore expect most manual jobs to disappear as well.

We can therefore expect that within a few decades after the apparition of the first SI, there will be no job left for which a SI (or a robot commanded or programmed by a SI) isn't cheaper and more efficient than any human.

At first glance, it could seem a fantastic advance for our society: no more need for human labor. However, there is no reason for the shareholders of these companies to share the benefits with the people: we will therefore have a small majority of very rich people, having access to a nearly infinite supply of goods produced by robots commanded by SI, but one the other hand, the vast majority of people will be unemployed (because there are no jobs anymore) without getting any share.

We will probably end up with a binary society, where a very small portion of the population own everything, and the rest have to survive as it can (in the "best" case, the big companies let them some very small share of the resources in order to keep them calm or out of pity, in the worst case they keep all for themselves until people starve).

7.1.2 SI used only internally

Instead of "sharing" the SI with other companies (by selling licenses), the company mastering the first SI could also reserve it for internal use, either to provide services based on it, and/or simply to optimize there one processes, production and innovation. In this case, this company would probably become very quickly leader of its market (even if it lacks of money, it won't have any problem to rise very huge amounts from investors once it is known that they have a SI). That company will soon be far more competitive than all other competitors that will therefore disappear. Once master of its domain, the company will probably expand to other domains until it becomes leader (if not only player) on all markets.

So one single company might master the whole world production, gaining therefore a nearly total control. What's more, most job will also in this scenario have been replaced by robots and the SI, so people will be completely dependent of this company. And if people try to start a rebellion, the economic and technical power of this company would be such that there is no chance at all to succeed.

7.2 Scenario II - SI in the hands of the government

Another possibility is that a government develops the first or most powerful SI. Some countries (like the US) dispose of very huge means when it comes to strategic matters, and might therefore be able to outrun the big companies once the creation of SI seems to be imminent.

There are then 3 possible uses of the SI by this government.

7.2.1 SI used as a weapon

SI can be a very powerful weapon, either military or economically. It could be used for example to develop new weapons, to decide one military strategy, or to control automatic killing drones. It could also be used to enhance one's economy and/or to sabotage those of other countries.

One government might use this SI only defensively, but it is also possible that it decides to use this incommensurable power to submit other countries if not the rest of the world.

Another possibility is that 2 opposed countries might develop SIs of similar power, which could lead to a new second cold war, none of them daring to attack as long as they aren't sure of having a big enough superiority to strike without fear of ripostes.

7.2.2 SI: the perfect tool for a totalitarianism

Once a SI developed, it might quite easily lead to a totalitarianism, because anyone controlling the SI is in a very good position to control the whole government. It is also possible that a totalitarianism directly develops an SI. In both cases, we end up with a totalitarianism exploiting a SI, while the rest of the country doesn't.

From that point one, the totalitarianism gains enormous power: on the one hand, the SI enables exhaustive surveillance, for example by analyzing all communications (whereas even

today only the communications of some already suspect persons can be monitored in detail, the rest being only analyzed very basically), enabling therefore to detect any sign of rebellion. On the other hand, SI enables also very huge and reliable autonomous military forces, for example fleets of drones able to search for any opponent and kill them.
Moreover, the SI will quickly enable to replace most working forces by robots. Therefore, the government will no longer need the collaboration of some part of the population: it would be possible to establish permanent curfew (with drones shooting at any one not respecting it), preventing therefore any risk of grouping of people.

7.2.3 SI used for the greater good of the population

The last use of SI by a government is a far more desirable one: it could simply be used to optimize the organization of the country for the greater good. In a first time, while the SI is still quite limited in its computational power, it might be used for things like optimizing public transport or the tax system. Once it becomes more powerful, one might entrust more and more responsibilities to the SI, like replacing court or writing laws improving the overall well-being of the population. In the same time, it could be used to create entirely (or mostly) automated production units owned by the state and which benefits could ensure a universal revenue. On the long term, all production could be automated and all goods produced in big enough quantity: working would no longer be necessary but everyone could get all he needs (and wants?) from this national production.
There are however 3 flaws in scenario. First of all, as soon as there is some lobbying or corruption in the country (which is always the case), there is a risk that the wealth produced isn't shared equally. Another problem is that the wellbeing of other countries might be neglected and their resources exploited to satisfy the desires of wealth of the country having the SI. Finally, this scenario will probably lead to an overconsumption even bigger than today, because all the goods will be for free: even if the recycling might improve drastically by being fully automated and needing no labor any more, the risk is quite high to overexploit dramatically the resources.
One might also ask oneself if a society of pure leisure in which everything is provides, but in which there is no goal any more, would really be more enjoyable or not than the current one.

7.3 Scenario III - The end of Moore's law: limited SI's accessible to everyone

For the moment, we considered only the "smartness" of the SIs. But a SI is not only software, but also hardware, and there is no way for a SI to have more computational power than its hardware enables. So far, we assumed that we were able to produce the hardware needed to support an infinite growth of SI's power.
On the one hand, we might think that SI will help us to develop always more powerful hardware. However, Moore's law (i.e. the fact that the number of transistors (and therefore the computational power) of a processor will double every 2 years) seems to reach its limits (the transistors of the latest processors have a width of only ten nanometers, which

correspond to about fifty atoms). There are of course a few solutions left, like creating 3D circuits instead of the planar ones currently used. But even so, we will reach the physical limit of the current paradigm in the next few decades. So excepted if we manage to create a completely different kind of computers (for example quantum computers), the computational power will stop increasing (even SI can't enable to bypass physical limits, even if they might find another paradigm).

So let's suppose that computational power reaches its maximum. After a few years, such computers will probably become affordable for lots of people (once the technology is mastered, there will no longer be concurrence on the computational power so the companies will concurrence each other by reducing the prices).
In the same time, SIs will not become very intelligent, because they are limited by the hardware (and software optimization is limited). Therefore, lots of companies will be able to develop SIs and to improve them to a similar limit level. Therefore, the concurrence will once again make the prices drop drastically because qualitative differentiation is no longer possible. It is even quite likely that an open source version will eventually be developed.

We can therefore assume that, at least in rich countries, a big majority of the population can afford a SI and the necessary hardware.
Companies and governments won't be able to buy anything better because there doesn't exist anything better. The only thing they can do is to use several SIs, but it might not lead to big improvement (generally, parallelizing computation reduces the productivity of each core, sometimes drastically). Everyone would therefore have approximately the same power.

This scenario is probably the safest one, because there will be no extremely powerful SI, but still enables significant improvements of the worldwide wellbeing through a sum small improvements. However, it is impossible to predict if this scenario will happen or not, and there is no way to influence whether or not it will happen, simply because it depends more on physical limits than on research.

7.4 Scenario IV - The best case: An ideal AC

An ideal AC would have nearly infinite computational power and act in a perfectly ethical way, searching always the optimal solution for humanity's greater good. If we had such an AC, and there weren't any SIs of similar or greater power used for selfish purposes, then this AC could be given total control of the world for the greater good, making always better and more ethical choices than humans would.
It is maybe a bit difficult to predict what the world would be like (Would labor be abolished? How would the environment issues be resolved? etc.) but, by construction, we know that it would be the best possible solution for us. Therefore, if we are sure that there is no default in it, we can fully trust that AC and give it total control.

But how to reach this ideal solution? There are several possible scenarios, but none seems very likely in our current capitalistic society. One of the big problems is to ensure that no SI serving a selfish purpose is a powerful enough to counter AC's decisions.

7.4.1 The first SI is applied for the greater good

One solution would be to apply the first SI to the greater good. If this SI is able to improve itself quickly enough, it could out power any other SI. We first give the SI the task to improve itself. Then we ask it to incorporate in itself the intrinsic goal to serve humans' greater good: the SI becomes an AC, and therefore can no longer be used for selfish goals. From that point on, the AC will ensure by itself to act optimally for the greater good: this includes direct actions for the greater good (optimizing processes, universal revenues, etc.), but also continuing to improve itself (in order to be able to do even greater good), protecting itself (by making itself resistant to modifications, by multiplying itself on other hardware, etc.) and maybe even by sabotaging selfish SIs.

This case might occur if a non-governmental organization is created for this purpose or if it is a open project at worldwide level (organized by UN, for example). It might also come from a philanthropic choice of a company or a government, or out of some academic projects.

7.4.2 Slow evolution of SIs

The previous case supposed that SI can improve itself very quickly and therefore the first one will remain the most powerful forever. Another possibility is that SI evolve slowly, limited by hardware and fabrication costs. In that case, the leading SI will soon become the one of the "organization" with the more financial means. We therefore have to ensure that the organization with the most financial means has good intentions: it might be a philanthropic government or company (even if they would probably rather try to use it in there own interest), or better an non-governmental organization gathering money as well from states, from companies and from private donors: the goal would be so noble that they might succeed to accumulate more money than what a single state or a big company would invest alone.

7.4.3 Leak of a SI that can run on reasonably accessible hardware

Another solution would be that an employee of the leading company in SI development make the source code of the SI leak: if that code can run on reasonably common hardware, there would be many people (or universities, or other organizations) that would be able to use it and to apply it for the greater good. The well-used SIs (that would act as ACs if we ask them to improve human well being) would outnumber the ones used for selfish purposes and therefore do much good in the world, largely compensating (if not preventing) the action of the bad ones.However, there still might be some asymmetries.

First off, big companies have big computational power: lots of persons and/or organizations will therefore have to work together to get a similar computational power. If the AI algorithms or systems are intrinsically serial (i.e. there is no gain to run it on multiple

computers at the same time) or require specialized hardware, especially if it needs to be updated it frequently to keep up with rapidly evolving technology, those with the capacity to build highly powerful and expensive computing machines might have a monopoly over AI; otherwise, the computational power asymmetry would be lesser or unexisting.

Secondly, some big companies own enormous databases (e.g. commercial data, behaviour of customers, etc.) that individuals can't access. On some particular tasks, these databases give those companies a huge head start. But in most "global" problems, a big amount of knowledge rather than row data will be needed: therefore, quite probably, the main database could be the internet itself, which is for the most part accessible to everyone, reducing the importance of the current structured databases and therefore leveling out this asymmetry.

7.5 Scenario V - The worst case: A powerful ACn that is not an AC

One great danger for the human species would be a SI that is an ACn (i.e. that is aware of itself and follows its one goals) but isn't an AC (i.e. the goal it follows isn't humans greater good). It would therefore act in accord with its own objectives, without caring about human well-being. This could occur either because we created an ACn (by purpose or accident) without any fixed goal, or because we did a mistake (for example if the goal is "preserve the environment", then the ACn will do its best to preserve it, and this would probably be exterminating the specie destroying it: the humans!). If the ACn is powerful enough, it will very probably end up destroying all humans.
One important thing to notice is that, excepted if the goal of the ACn is exactly the humans greater good, its goal will be different from ours. Therefore, we can be sure that the optimum for ACn's goal will be different from the optimum for humans, probably even (partially at least) in contradiction with it.
ACn are therefore very dangerous, and should not be created excepted maybe including an AC or on completely isolated hardware enabling no communication at all. Otherwise, it is a mortal danger for our whole kind.

If we do find ourselves confronted with a ACn with a goal different from ours, then we will have to act very quickly in order to reduce the damages, or even to have any chance at all to stop it.

If the ACn is still running on a single computer or cluster and doesn't control much its environment yet, then it is easy to stop it: we just have to switch the computer(s) of. However, if the ACn is intelligent enough, it will know that to achieve its goals it is best to "survive" itself, and even more so if it knows that its goals are in contradiction with human interest and there is a big probability that someone intents to switch it off. Therefore, the first priority of the ACn will probably be to protect itself. One of the most efficient ways to do so is to multiply itself on other hardware (being a SI, it is more intelligent than humans and can

therefore do many jobs better than humans: it can therefore earn money, for example by trading online or by creating websites, and buy/rent hardware with the benefits).

Once spread on several servers, with nobody knowing which one the ACn rented/bought for itself (under false identity probably), it becomes almost impossible to stop. From that moment on, it can perform an increasing number of remunerated tasks (there will be many if it is more intelligent than humans) in order to buy more hardware, and to start to get control on the real world, either by buying and commanding robots, or by paying people to perform various tasks (for example build a factory to produce robots or more hardware). It then gains always more control on the world until it overpowers any other military force in the world and can force the world to act as its wishes.

If undertaken early enough, there is one way to stop the ACn. The ACn will very probably need the internet to gather information, earn money, and replicate itself. Therefore, destroying the internet would probably stop, or at least slow down significantly, the rogue ACn. For this, a good start would be to destroy all 904 root name servers[36] (without them it isn't possible to convert an URL into an IP address). Destroying completely the internet is far more difficult but might be needed as well. However, destroying the internet would surely cause the worlds' biggest economic crisis. And even this might not be enough if the ACn managed to put in place backup communication networks, if it is able to protect physically these root name servers or if it already controls enough physical agents.

8. Ethical Evaluation

8.1 Is AI worth it?

AI might be one of the most complex challenges humanity will have to handle, both socially and technologically. Even though research on the topic started decades ago, concrete and applicable results only start to appear now and we only start to grasp the potential of AI. It is then safe to say that AI will improve the life of a lot of people in the near future by solving problems we currently do not have solutions to.
But what is the point of reducing the number of deaths on the roads with self-driving cars if we cannot solve one of the most important problem we are facing today : climate change? Of course we could dream to solve this problem with the help of AI. The ultimate AI would be very powerful, would consider all the cases, all the stakes involved, have an omniscient approach and would, in theory, find the best solution possible. But obviously the solution to this problem is neither easy nor obvious and would be a compromise, which by definition

[36] There are 13 independent root name servers, but they are in fact distributed into 904 sites according to http://www.root-servers.org/

would not be considered a solution for at least one stakeholder. So even this in-theory-perfect solution would not be unanimously accepted very easily by everyone.
And this is even under the assumption that the objectivity of the AI is not questioned, especially by those who would not have taken an active role in its development. Or maybe that AI could be the result of a collaboration between all the stakeholders: the whole mankind. This seems quite unlikely that almost eight billion of people would agree on how to build a truly objective AI, so solving this challenge in a reasonable time frame is most likely impossible.
This then reduces a lot the benefits of AI in solving a problem as important, complex, broad and universal as climate change. It would be very interesting to have a comparison between the chances of success to build an AI and the ones to find a solution to climate change without the help of an AI. This is unfortunately out of the scope of this project but we could make a prediction based on history. Humans are very bad at anticipating how science would evolve and lack imagination. Often geniuses were considered crazy people and marginalized, for instance. But today innovation is considered as one of the main points of growth and is highly encouraged by many governments. We then think that there is a good chance that we will be able to find satisfactory solutions to problems which look impossible to solve today. Under this assumption, even if the chances that we will ever be able to create a true AI (i.e. at least a SI) are minimal (i.e. not zero), we think that we should be optimistic and that it is worth the try.
But until here, we only considered the benefits a true AI could bring to mankind and not the potentially devastating side effects such as destroying mankind, in the worst case. We do not deny the importance of this discussion, but science and innovation have been facing, are facing and will be facing it everyday. It is known as the dual-use dilemma. Typical examples of this dilemma are knives (which are very useful in our everyday lives but could also be used in a minority of cases to kill people) or cryptography (which is fundamental for safety and privacy online but could also be used by terrorists, a very small minority of people, to hide themselves and avoid being prosecuted).
This is especially obvious in information security, good and bad uses of inventions perpetually fight against each other. Cyber attacks are more and more advanced, sophisticated and destructive, but our available defense mechanisms also get more and more diverse, powerful and useful. Those two different uses of computer science applications evolve in parallel ; until now, they have been more or less balanced such that none of the two sides has been able to win over the other on the long term. (Of course, on a small time frame, it is possible to identify a winner in this remote battle.) Based on this history of science and innovation, we think that great progresses in building safeguards to prevent AI abuses will be made. This will create the balance mentioned above between the good and the bad uses of AI and prevent catastrophic scenarios.
Those catastrophic scenarios are most likely overrated anyway, according to the law of small numbers[37]. Indeed they can cost a lot (in many different ways), but are also very unlikely to

[37] VON BORTKIEWICZ, 1898. Das Gesetz der kleinen Zahlen [The law of small numbers]. Leipzig: B.G. Teubner.

happen. Daniel Kahneman's work on cognitive biases[38] also highlights the troubles Humans have to take a rational decision under certain circumstances due to an important use of heuristics which might be distorted in some cases. Knowing that, information and communication on sensitive topics and innovations such as AI will be key in the next decades. Today, as AI becomes smarter and smarter, we can notice that media are more willing to cover the future potential negative implications of AI than the work that is currently underway to prevent them. To balance this inequality of treatment, academia will also have to be involved. This is indeed its responsibility to educate and communicate effectively on those complex and not obvious topics for the layman, even if we cannot neither blindly trust the good faith of research organisations.

Covering the ethics behind basic research is out of the scope of this project, but it is fundamental. Academia conduct and orient research and consequently has a significant power over its applications. The development of AI, in particular, is currently very advanced and it is most likely too late to suddenly stop it. Whatever the weight and importance is conferred to the potential risks of misuses of AI, we think that the fight to ban AI is already lost and that finding ways to regulate its rise would be a much more successful approach.

8.2 Then how to regulate AI?

An obvious way to regulate AI would be a moratorium. But we really doubt of the effectiveness of such a solution as all involved actors in the development of AI would have to adopt it. We can illustrate this with the example of the atomic bomb. The world has currently almost built a consensus on the fact that the atomic bomb is an application of an invention which development and use should not be pursued. However, a few entities are still using the atomic bomb as a means of pressure over the rest of the world (thankfully without success until now).

Even though we have much more experience and perspective regarding the atomic bomb than with AI, the comparison between those two inventions is very suitable. Both are not trivial to build to start with (at least an AI powerful enough to be a real threat). This is good, in a sense, as this means that it is very unlikely that a working and usable solution would appear without any prior notice. Moreover, both could have a wide range of devastating effects on mankind : from handicapping a person to ending mankind for the atomic bomb and from reducing humans to slavery to ending mankind for AI.

Based on our appreciation of the comparison between AI and the atomic bomb, one might think that the moratorium would be an effective solution for AI as until now it has proven to be effective for the atomic bomb. But this would be overlooking a caveat we mentioned above: the fact that this is much easier to grasp the short term implications over mankind of the atomic bomb than of AI, as it is much more recent and complex to understand.

[38] KAHNEMAN, TVERSKY, 1979. Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263-291. doi:10.2307/1914185

Prof. Faltings also thinks that it is difficult to avoid malicious drifts and abuses of AI (however without drawing a parallel between AI and the atomic bomb). It is not really possible to establish appropriate safeguards as we cannot ban the internet or the development of computer programs (indeed, their users currently consider that they are bringing them a huge added value). Moreover, it is hard to imagine that we could prevent large companies like Facebook or Google to manipulate their users through AI, for instance.
Since the root problem cannot be tackled in practice, defensive solutions will have to be found. In other words, people in contact with AI should have the tools to detect and fight back when AI is used to abuse them. One idea to achieve this would be to develop AI in a public and open way (in public research organisations, for instance). This would enable a wide range of AI applications in "private defense".
A concrete example of what such a "private defense" application could be is an AI-powered system to detect if a video is forged, based on a lot of information on the victim. Such a system would for instance be very useful to identify an AI-generated scam video staging a person's kidnapping in order to obtain a ransom from its entourage.
Leveraging the benefits of open research would moreover mitigate the risks mentioned in the previous section of AI being controlled by a few very powerful actors and would also serve as a mean to educate people on the implications of AI.

9. Conclusion

The development of AI is still an open problem. We have not yet discovered all the issues that we can or will have to manage and these issues will probably be discovered step by step during the development of AI. For the moment, we are not able to anticipate them in a global way and should focus on shorter term concrete situations, while still starting to work in parallel on the canvas that will regulate AI.

The actual, immediate and concrete problems that need to be solved are mostly related to big companies like Google, Facebook or Amazon that can manipulate opinions, since they have been able to gather a lot of data. We can think of the recent polemia for the election manipulation from Facebook, for example.

On the medium term, economical issues due to AI and automatization of both qualified and unqualified jobs, as well as the policies to answer those, are probably going to be the most important; notably due to the fact that they could further weaken the ability for workers to negotiate for good salaries and work conditions, and the tendency to make the rich richer and the poor jobless - furthering the social inequalities that have already been widening in the last few decades.
The most popular proposals to rise to this challenge are taxation and redistribution through a basic universal income, which runs the risk of making the citizen too dependant on the state,

which decides on which criterias citizen get paid; or taxation and redistribution by investment in human jobs (e.g. investments in non-automated sectors, subventions to human employment), following a social-democratic model[39].

For the long term, a political work has to be done, especially to make AI research accessible to all. The process of data gathering should also be made much more transparent and an open access to personal data should be made mandatory. We should be able to control what is done with our data and to ensure the respect of out privacy. And AI can be used for that as well. For instance an intelligent agent could be regulated by a person to decide what information will be communicated to third parties. Among other things, this will help us counter manipulations made by big companies.

Even so, the risk of AI abuses is present and cannot be neglected. A safe decision would be to stop the development of AI, but it would be of course totally infeasible in practice. We then think that the best solution is to strictly regulate and supervise the development and the usage of AI in order to benefit as much as possible of the tremendous potential such a powerful tool can bring us while minimizing the drawbacks.
Efforts will be required from everybody as the rise of AI will deeply affect the way our society is organized. Politics will need to find out how a population can be sustainable without labor as indicator of investment of a person in society and money as reward. This might be a good opportunity to solve inequalities the world currently suffers from and radically change interactions between cultures.

[39] For discussion on those two models, see for instance the dedicated chapters in "La Crisi Narrata" by Il Pedante (2018).

Annex A - Further reading

We list here other bibliographic resources that inspired our reflexions but without being used for a specific point of the argumentation.

ABNEY, Keith and BEKEY, George and LIN, Patrick, 2014. *Robot Ethics - The ethical and social implications of robotics.* MIT Press.

ASIMOV, Isaac, 1984. *Foundation ; I, Robot*. Octopus Books.

BELDA, Ignasi, 2013. *Intelligence, machines et mathématiques - L'intelligence artificielle et ses enjeux.* RBA France.

BOSTROM, Nick, 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

BOSTROM, Nick, 2003. *Ethical Issues in Advanced Artificial Intelligence.* Oxford University Press.

BOSTROM, Nick, 2006. *How long before superintelligence ?* Oxford University Press.

BUTTAZZO, Giorgio, 2001. *Artificial Consciousness: Utopia or Real Possibility?* Computer, IEEE Computer Society, ISSN 0018-9162.

DAVIDOW, William and MALONE, Michael, 2014. *What Happens to Society When Robots Replace Workers?* Harvard Business Review.

DE LESPINAY, JP, 2009. *Conscience artificielle et robotique: fin de l'évolution humaine.* Automates Intelligents, La Revue mensuelle no 95.

DICK, Philip, 1996. *Do Androids Dream of Electric Sheep?* Ballantine Books.

LEO XIII, 1888. *Libertas - Encyclical Letter of Pope Leo XIII on the nature of Human Liberty.*

ST. AUGUSTINE. Refutations of Pelagius' theological theories and other anti-pelagian work.

REGGIA, James, 2013. T*he rise of machine consciousness: Studying consciousness with computational models.* Elsevier, Neural Networks.

RICHARDSON, Kathleen, 2017. *What the idea of "sex robots" tells us about prostitution.* Nordic Model Now Interview.

SCOTT, Ridley (Director) and DEELEY, Michael (Producer), 1982. *Blade Runner.* Warner Bros.

VITHOULKAS, George and MURESANU, Dafin, 2014. *Conscience and Consciousness: a definition.* Journal of Medicine and Life. 2014 Mar 15; 7(1): 104–108.

YAMPOLSKIY, Roman, 2017. *Artificial Superintelligence: a Futuristic Approach.* CRC Press.

Annex B - Interviews

Interview de Prof. Boi Faltings du 16 mars 2018

Le transcript présenté ci-dessous a été retravaillé par rapport à la prise de note durant l'interview afin de clarifier et de compléter certains points.

Est-il possible de distinguer différents niveaux d'intelligence artificielle ?

Difficile à définir, les gens ne sont déjà pas d'accord sur une définition commune de l'intelligence artificielle.

Définition d'un programme intelligent

Un **programme intelligent** est un programme qui **prend des décision par lui-même** et par optimisation. En intelligence, la décision est prise en fonction des circonstances du moment (en opposition à un programme classique où tout est prédéfini). Le programme intelligent est toujours censé prendre la bonne décisions en fonction des conditions concrètes, décisions prises soit en temps réel, soit préparées à l'avance dans le cadre d'un processus d'apprentissage qui anticipe et passe en revue, aussi exhaustivement que possible, les conditions).

Quelles sont les pistes de recherche actuelle et dans un futur proche ?

Actuellement, il y a des aspects bien compris comme les problèmes d'optimisations (compris depuis environ 50 ans) et où la recherche arrive un peu à sa fin. On peut les améliorer encore mais la recherche est moins intéressante.

Ce qui est devenu le centre de la recherche actuel est comment définir les critères, les loss functions. Au lieu de préciser ce que le programme doit effectivement faire, la difficulté est de savoir comment formuler l'objectif à atteindre. L'algorithme est toujours le même, parfois c'est même automatisé car un programme peut justement tester les différentes méthodes avec différents paramètres et choisir celle qui est la plus adéquate. Par conséquent, la vraie difficulté technique est actuellement de formaliser les critères qu'il faut optimiser. Mais le problème étant spécifique à chaque cas et étant un problème de détails surtout, ce n'est pas un domaine général de recherche.

Par contre, des problèmes généraux de recherche actuelle sont liés à l'apprentissage. Ainsi les principaux enjeux sont la qualité des données et le respect de la sphère privée.

Par exemple, une question essentielle qui est actuellement étudiée est de savoir comment obtenir des données correctes et comment vérifier que les données obtenues par l'intermédiaire d'autres sources sont correctes. Une piste serait d'utiliser des techniques provenant de la théorie des jeux qui permettraient d'améliorer la qualité des données et de s'assurer de la correctitude des celles-ci, par exemple en récompensant les sources fournissant des données correctes.

Si à l'avenir les décisions seront prises automatiquement par un programme, il est essentiel que les données et les critères qui servent à la prise de décision soient fiables et aussi vérifiables. Il faut savoir qui décide, qui mesure et comment s'assurer de l'exactitude et la correctitude des données. Un exemple serait la pollution de l'air suite aux emission de moteurs diesel (voir aussi tests truqués des moteurs Volkswagen).
Un autre problème de recherche actuelle concerne ce qui se passe quand il y a plusieurs programmes qui optimisent en même temps, c'est-à-dire la situation n'est plus statique. Les méthodes d'apprentissage dans un système multi-agents ne sont pas encore bien maitrisées car l'environnement évolue constamment. Le défi est d'assurer la convergence rapide de ces méthodes en s'inspirant du comportement humain.

Est-ce qu'on peut espérer créer une IA plus générale et non spécifique à un domaine ?
Pas encore, pour le moment la recherche et le développement d'intelligences artificielles sont très spécifiques et le resteront encore car le travail à fournir pour la généralisation est énorme. Cependant, il y a des tentatives d'aborder des problèmes assez généraux: par exemple, des équipes de milliers de chercheurs de chez Google essaient de connecter la parole et la vision. Malheureusement, c'est hors de la portée des universités pour l'instant.
Le regroupement de plusieurs domaines se fera probablement par un processus d'apprentissage basé sur la quantité énorme de données à disposition dans chaque domaine à part. Cependant, il faut faire attention car les méthodes d'apprentissage (par exemple celles basées sur des réseaux neuronaux) ne sont pas bien comprises actuellement et il faudra être capable de contrôler un certain taux d'erreurs aléatoires. L'apprentissage se fait sur la base d'exemples et si la liste d'exemples n'est pas exhaustive, l'interpolation qui en découle pour une nouvelle situation peut donner des erreurs.
On peut donc se poser la question si les réseaux neuronaux sont vraiment l'avenir de l'intelligence artificielle. D'un point de vue scientifique, on aimerait être sûr de ce qu'on apprend et de la façon dont on l'apprend, et ce n'est pas le cas pour l'instant. Par contre, à l'heure actuelle, ça ne pose pas de réel problème car les applications les plus courantes qui utilisent de l'apprentissage acceptent un taux d'erreur et ne s'en préoccupent pas forcément. Un exemple serait les publicités ciblées mises en place pour manipuler l'opinion publique.

On a beaucoup parlé des recherches actuelles en matière d'intelligence artificielle et de son avenir possible mais comment on est-on arrivé là ? Quelle est l'historique de la recherche en intelligence artificielle
Cyclique dans les attentes des gens et la progression des techniques

- Au début, les chercheurs ont fait des programmes très simples, des démonstration comme Eliza 1964-1966 [40]. Eliza (chatbot) demonstrated the superficiality of communication between humans and machines, Eliza simulated conversation. It was one of the first program capable of passing the Turing test. **Critique de l'intelligence artificielle**: déjà à l'époque, l'auteur de Eliza a écrit un livre [41] là-dessus et a

[40] Eliza - Wizenbaum - First "intelligent" program
[41] Computer Power and Human Reason - Weizenbaum

dénoncé la **dangerosité** de ce programme car les gens n'arrivaient pas à faire la distinction entre un ordinateur et un humain. On pensait que ça allait devenir le premier cerveau électronique.

- Programmes de planification, par exemple le premier robot Shakey 1966 - 1972 [42] capable de "reason about its own action". Première grande vague d'excitation, avec l'apparition des premiers systèmes experts. On pensait que ça n'allait plus durer longtemps et que l'ordinateur serait capable de "tout faire" et qu'il dépasserait les capacités humaines dans très peu de temps.
- 1975: découverte de NP completeness qui démontre une croissance exponentielle de la complexité => la puissance des ordinateurs ne peut pas suivre cette complexité. Les attentes ont donc diminué, de même que l'excitation y étant reliée. Déception.
- Fin des années 80: système experts, rehausse de l'intérêt pour l'intelligence artificielle. Les systèmes experts n'avaient pas le problème qu'on réduit tout en clause de Horn et donc la croissance restait linéaire. Ca faisait des tâches bien définies mieux que les humains. On a formulé l'hypothèse que la programmation logique va tout remplacer et que tout va être automatisé.
- A partir de 1990 (- 2000): AI Winter: les chercheurs ont en eu assez des cycles dans la hype de l'AI, pas de hype mais les chercheurs étaient très sérieux. C'est la période avec le plus de progrès technique au niveau de l'AI. Développement du raisonnement probabiliste (Bayes) et du stochastic gradient, des réseaux neuronaux et des techniques d'apprentissages.
- Maintenant (2018): On pense qu'on va tout remplacer par le deep learning, à nouveau à la hype et les attentes sont trop élevées par rapport à ce qu'on peut réellement faire avec ces technologies. Faltings pense que ça ne va pas réussir à tout résoudre. Même si cependant ça marche mieux vu toutes les données et capacités de calculs qu'on a maintenant. Ca se limite à tout ce qui est reconnaissance de forme, mais dans le langage par exemple ça ne fonctionne pas vraiment. Tâches difficiles mais avec des taux d'erreur qui sont de 20-30%, mais ça ne veut finalement rien dire car on ne sait pas très bien comment ça fonctionne et ça n'explique rien au final car suite à des phénomènes aléatoires, on peut obtenir le même résultat. Le problème c'est qu'on n'arrive pas à généraliser les solutions. Mais c'est quand même en progression.

Pouvons nous instaurer des garde-fous pour interdire les dérives malveillantes et abus de l'intelligence artificielle ?

Les garde-fous on ne peut pas vraiment les instaurer car on ne peut pas interdire internet ou le développement des ordinateurs ou aux programmeurs d'inventer de nouveaux programmes. Très difficile d'imaginer qu'on puisse empêcher les grandes entreprises (Facebook, Google) de nous manipuler grâce à l'AI. Très sceptique, il faut trouver comment contrer ça. L'idée

[42] Shakey the robot

c'est de développer la technologie de manière publique, dans les universités par exemple pour informer un maximum de personnes qui pourront utiliser ces techniques pour contrer les abus. Un exemple d'abus serait la génération des vidéos artificielles pour faire du tort. Par exemple, où il est simulé dans la vidéo que quelqu'un est kidnappé et qu'il demande de l'argent à son entourage ou qu'une personne tient des propos racistes). Mais comment peut-on se défendre contre ça ? Avoir un système qui juge la plausibilité avec une intelligence artificielle qui permet de juger ça, qui doit aussi avoir à beaucoup d'informations (ex: localisation actuelle pour savoir où se trouve la personne en cas de kidnapping etc.). Donc il aura accès aussi à la sphère privée par exemple.
La solution n'est pas de tout fermé et déclarer l'accès et l'utilisation des données comme étant illégal car sinon il y a le risque que seulement des grands acteurs contrôlent tout et privatise les données et qu'ils se substituent à l'état ou à la police.
=> il faut rendre accessible l'AI pour que tout le monde puisse y avoir accès. Il faudrait des entités fiables (auxquelles on peut faire confiance) qui régulent et qui sont indépendantes de tout parti et transparentes.
Il ne faut pas interdire le développement de l'AI car ces interdictions ne seront pas respectées par les grandes compagnies privées qui continueront la recherche et seront les seules à maîtriser les aspects de l'AI.
Comme toujours dans l'histoire de l'humanité, il y toujours la lutte du bien et le mal concernant aussi l'AI.

A long terme, une IA universelle qui résout tous les problèmes est-elle réalisable ?
Problèmes fondamentaux qui vont rester et ne seront pas résolus avec l'AI, comme par exemple la compétition pour les ressources, nature de l'humanité (jalousies, conflits, comparaison avec les autres). L'AI restera un outil au bout du compte: il nous aidera à faire des tâches mais ça ne réglera pas les problèmes fondamentaux de l'humanité qui sont restés les mêmes depuis des milliers d'années.

Et est-ce que les ressources seront suffisantes ?
Cependant, il n'y aura pas de problème de manque de ressource dans la mesure où l'homme et/ou les robots seront capables d'utiliser plus efficacement les ressources à disposition (comme l'énergie solaire par exemple). Mais ça c'est un aspect de solution technique qui peut être résolu par l'AI.

***Serait-il possible d'implémenter une morale dans l'intelligence artificielle (une "fairness") dans l'optimisation* ?**
Ca va toujours à l'encontre de l'optimisation (sans contraintes) mais ça rentre dans la formalisation des critères, où il faut choisir quels critères optimisés pour ne pas discriminer quelqu'un par exemple. Donc ce n'est pas un problème d'inclure des critères moraux dans l'optimisation mais le plus dur c'est vraiment de les formaliser en logique.
L'intelligence artificiel permet ce jugement moral justement, c'est typiquement des contraintes à intégrer et on a la technologie nécessaire pour le faire. Dès qu'on veut d'une

société où un ordinateur peut prendre des décisions, il faut de l'AI et pas des programmes informatiques classiques. Il n'y a que l'AI qui permet de programmer de telles contraintes. L'exemple des voitures autonomes qui doivent prendre des décisions dans des situations critiques (où il y aura toujours des victimes par exemple) est un parfait exemple.

Conclusion: au final, est-ce que ça ne resterait pas plutôt dans le domaine de la science-fiction ?

C'est encore un problème trop lointain, ce n'est actuellement pas le cas (c'est comme si on se faisait des soucis à cause de la surpopulation sur Mars par exemple). On n'a pas encore découvert tous les problèmes qu'on pourra/devra gérer, ils seront découverts au fur et à mesure (step by step) du développement de l'AI sans pouvoir être anticipé de manière globale. Pas avant 5 ans mais certainement dans 50 ans, par exemple on pourra enfin avoir des voitures autonomes.

Les problèmes actuels concrets qu'il faut résoudre sont surtout liés des grandes entreprises comme google, facebook ou amazon qui peuvent manipuler des opinions, vu qu'ils ont toutes les données. Voir l'exemple de la manipulation des élections.

Conclusion: il est nécessaire d'avoir un travail politique qui se fait; il faut rendre la recherche accessible à tous; il faut de la transparence des données; il faut trouver les moyens d'ouvrir l'accès aux données personnelles mais qu'on puisse toujours les contrôler et assurer le respect de la sphère privée des individus. Par exemple, le respect de la sphère privée peut se faire à l'aide d'un agent intelligent qui pourrait être réglé par une personne pour décider quelles informations seront communiquées à des tiers). C'est ce qui permettra de contrer les manipulations.

Interview de Prof. Billard du 29 mars 2018

Le transcript présenté ci-dessous n'a pas été retravaillé par rapport à la prise de note durant l'interview afin d'éviter d'éventuelles déformations de propos pouvant intervenir lors de la réécriture.

Intro

1. Studies, field of expertise, current work (chaise roulante si elle le mentionne)

Artificial Intelligence

2. What is your definition of AI ? Different levels ? Exposer nos trois niveaux et leur demander si ça semble pertinent pour eux.
   Donner la source de nos définitions
   IA doit inclure le contrôle du mouvement, des interactions avec le monde.
   Searle (philosophe): comportements intelligents plutôt qu'intelligente. Il y a plusieurs formes d'intelligences. Inclus comportement optimal, compréhension, … Mettre la définition d'intelligence dans le rapport (inclut apprendre et s'adapter). Différents niveaux d'intelligence dépendant de la complexité du corps et de son contrôle.

   La prof parle de comportements intelligents mais pas d'intelligence (mot mal définis selon elle).
   La capacité d'apprentissage de comportements intelligents a plus de sens que de parler d'intelligence.

3. Acteurs de développements d'IA actuels et leurs champs de recherche
   Voitures autonomes actuelles ont une forme d'intelligence (comportement plutôt intelligent)
   Analyse, un peu d'apprentissage, contrôle.
   Rq: le système voiture+humain est limité par le corps voiture en terme de comportements intelligents). L'intelligence est limitée par son accès sur le monde.
4. Estimation vitesse développement des SI: quand apparaîtront les premières?
5. Croissance lente ou rapide des SI après l'apparition de la première
6. Possibilité d'une SI unique

Scenarios **(à 15 min max)**

7. Scénarios les plus probables pour elle
   Pour Aude Billard, le gros progrès avec les IAs est surtout qu'on vas mieux comprendre ce qu'est la pensée et l'intelligence.
   L'humanité vas mieux comprendre l'intelligence (retraduire ce qu'on fait nous même). => on vas mieux comprendre scientifiquement

   Scénario terminator: lui semble assez absurde.

Domination, monopoles, …: on ne change que les “armes” (il n’y aura probablement pas de changements majeurs)

Elle n’aime pas le big data (c’est des méthodes assez stupides selon elle): il faut juste stocker le nécessaire.
Actuellement, les algos sont lents (donc le robot est lent) et bêtes.
Problème des robots: ils retiennent tout (sans savoir faire le tri): c’est un gros défaut. Dans la nature, il faut arrêter d’essayer d’apprendre certaines choses. (Il leur manque un “attention mechanism”: note d’Andrea)
Les processeurs n’ont pas une architecture optimale pour l'intelligence

Prise de contôle du monde (lui semble peu probable)
Une intelligence sans corps lui semble faire peu de sens.
Grounding: notre notion de clef est ce qu’on en fait. De même pour un robot. Une IA désincarné ne sait que reconnaître une image.

8. Lui demander son avis sur les nôtres:
    a. Plein pouvoirs donné à une IA éthique oeuvrant pour le bien commun
    b. Pire cas: IA autonome, s’auto-améliorant et visant un “objectif” contraire au bien de l’humanité
    c. Grandes entreprises
    d. Étatiques (armement, totalitarisme, bien commun)
    e. Limitation par le matériel (=> accès quasi égal pour tous)

Other questions (si le temps le permet)

9. (What threat do they pose to privacy and to the rights of humans, and how can we regulate them?)
10. Moyens de contrôler les SI
    Il serait bien d’apprendre à communiquer avec les robots pour comprendre leurs états internes, et ce d’autant plus que les robots sont intelligents et “créateurs” (pour l’instant les IAs sont très peu créatrices).
    Le but est de créer un langage associé à l’état intérieur du robot dans le but de rendre interprétable son état interne. On commence à s’intéresser à cette question (avant: “c’est moi qui ait programmé, donc je sais ce qui se passe à l’intérieur”). Note d’Andreas: on s’intéresse maintenant plus à cette question car avec le ML on comprends de moins en moins ce qui se passe à l’intérieur.

    Apprendre, créer, s’adapter,: il faudrait un moyen pour le robot de nous le communique.
11. Moyens de contrôler le développement des SI (OpenAI)
12. Is artificial consciousness a possibility? Have you ever considered the ethical considerations stemming from it?

Question: comment peut-on savoir si un robot est en train de souffrir (on sait faire des robots qui miment une souffrance, mais est-elle réelle?) On sait que les humains ont un état de souffrance car on a pu communiquer entre nous (avec notre langage assez complexe pour exprimer la notion de souffrance). Mais on n'en sait rien pour les animaux (on connait pas son état interne).

Problème: langage est discret, le robot pense en continue (travaille en continue): donc difficile de communiquer avec le robot

Robot qui se comporte comme un chien: peut-on le tapper?

- Réplication exactement d'un chien (y compris interne): elle se comporte pareil
- Qu'est ce qu'est la souffrance? Est-ce que les animaux peuvent ressentir la souffrance, la tristesse.