Artificial Intelligence. Is the Power Matched with Responsibility?

Marija Slavkovik

1 Introduction

It is safe to assume that most people today have heard of Artificial Intelligence (AI) and at least half of them are of the opinion that this technology will change their lives and needs to be carefully managed (Zhang and Dafoe, 2020; Cave et al., 2019). For some, the myth of AI and the science of AI are hard to separate. This chapter looks into the state of the art of AI as a science and critically assesses the threat to society posed by how AI is used and explored today. The main position of the chapter is that the highest threats that AI imposes comes from our miss-attributing human abilities to computers and replacing human cognitive abilities with computer programs without fully understanding the limits of either.

AI tickles the fantasy of humanity both as the great promise of doing away with drudgery and as the great threat of creating the agents of our own demise, ranging from autonomous killer machines to mechanical slaves that direct the wealth of the world away from the common people into the hands of a small elite. To safely use AI in the future we must understand the tool that AI is today and how to best use it. We hope to start this process with this article that relies on the foundations of AI rather than on the speculation of what AI can be.

The research field of AI can be traced back to a workshop at Dartmouth College in 1956 (Moor, 2006a). The workshop was held under “the conjecture that every aspect of learning or any feature of intelligence can be in principle so precisely described that a machine can be made to simulate it” (Moor, 2006a). AI today encompasses many subfields, the precise mapping of which is a task that itself requires automation (de Kleijn, 2018). In Sections 2 and 3 we use the automation of reasoning as an anchor to describe what the AI subfields are. In Sections 4, 5 and 6 we discuss the limits and related threats of AI as we know them today reflected in its subfields. Often the main threat of AI is seen as insufficient human control in AI operations. In Section 7 we address specifically the challenges of the human supervisory role in AI. Lastly we summarise the frequently repeated concerns and fears about AI and outline suggestions for how the threats of AI overall can be mitigated.

2 What is AI?

The main existential threat of AI is in attributing to it abilities that it does not have. Therefore the first line of protection is to understand what AI is, and what it can do. In this AI introduction we use simple examples that everyone can understand, rather than reports on AI applications as the reader is perhaps accustomed to seeing. Popular science writing on AI focuses on examples of the best and the worst applications, with the threat of AI being discussed around these examples. We list some of the reasons why this approach can be very misleading.

Popular science rarely reports on research breakthroughs, for the simple reason that it takes a long time for a research breakthrough to reveal itself as such. As a consequence we are led into the impression that the cutting edge AI is about predicting the future and targeted advertising. Considering finished applications when evaluating the AI impact leaves society with two choices regarding AI: take it as it is, or leave it all together. However AI methods and artefacts are made by people and can be altered by people at any point of deployment. Looking at finished applications, it is easy to disregard this fact and feel powerless at the mercy of some “corporate overlords”. We can do much with early education to accomplish human-empowering AI and this opportunity should not be disregarded.

A single definition of AI is not agreed upon. “Artificial Intelligence” was coined by John McCarthy to distinguish the field from cybernetics, but he himself recognised (Mitchell, 2019) that the name is not the most adequate. The core of the issue with the name is that it is tying the discipline to a concept that itself is not well understood, namely “intelligence”. Definitions of AI often risk being too narrow, alienating some aspects of the goals, methods, and purposes of AI. For the purposes of this chapter we select two definitions that most clearly highlight the impact AI has on the society in which it is developed and deployed.

Bellman (1978) defined AI as the automation of activities that we associate with human thinking, i.e., cognitive activities. This definition circumvents issues of whether what a machine does can be eventually considered thinking in the same sense of the word as when it is applied to people. The question of “can computers think?” is not the primary focus of this chapter, since thinking is not a necessary feature that computers need to be able to cause harm. Therefore this definition serves us well. Furthermore, it is the application of AI in the role of replacing human activities that has been recognised as the main source of concern (Kearns and Roth, 2019; Dignum, 2019; Mitchell, 2019; Noble, 2018; O’Neil, 2016; Algorithmic Watch, 2020).

For Poole and Mackworth (2017) “Artificial intelligence, or AI, is the field that studies the synthesis and analysis of computational agents that act intelligently.” An agent here is an entity that acts in an environment, whereas a computational agent is “an agent whose decisions about its actions can be explained in terms of computation.” A lot of the recent progress in AI applications have been done not by considering a “whole” agent but by building programs that replicate certain aspects of agency such as decision-making, reasoning, and situational “awareness”. However, it is this separation of cognitive processes from the agent that we argue is important to identify because it directly highlights the often invisible shortcomings of AI.

Of all the cognitive processes, reasoning is perhaps intuitively most closely related to intelligence. Reasoning is the process of considering something in a systematic and logical way. Namely, we are doing reasoning when we are considering events together with rules collected from previous experience to evaluate observed events (see Figure 1 for an illustration). By doing reasoning an agent is, in a way, organising the information it has about its environment, in order to use it more effectively when pursuing its goals.

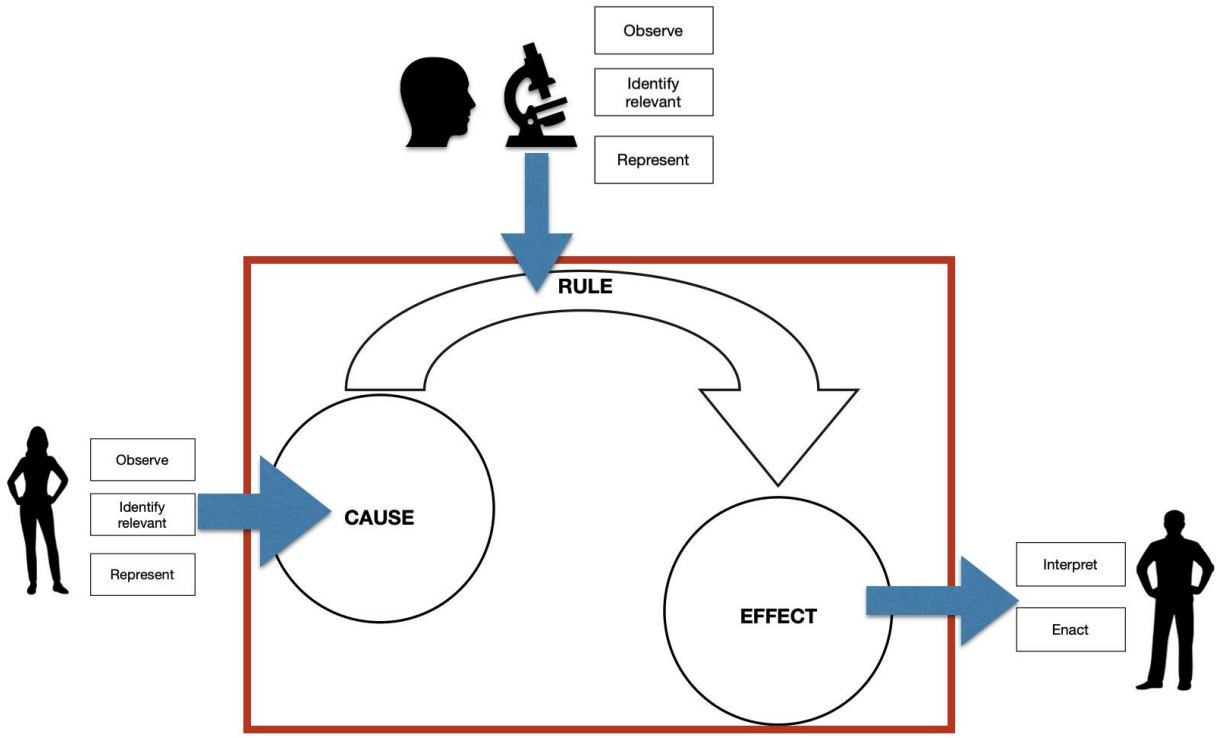

Figure 1: Logic reasoning can be seen as the process of distinguishing observations as cause and effect connected with some rule. File: reasoning1.eps

We can distinguish at least three types of logic reasoning that AI is trying to replicate computationally: deduction, abduction, and induction.

Deductive reasoning, deduction or deductive inference is the process of reasoning where we use a set of observations and rules of implication (or derivation) to reach a logically certain conclusion (see for example (Russell and Norvig, 2015, Chapter 9)). Using Figure 1, deductive reasoning is what we do to find out the "effects" when we know the related causes and rules. To experience deductive reasoning one can try to solve the following puzzle (Kleiner, 2009): Jack is looking at Anne, but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person? The solution of the puzzle is one of the answers: yes, no, or cannot be determined[1].

Abductive reasoning, abduction or abductive inference is the process where given a set of observations and rules of inference (implication or derivation) we are looking to find the best explanation for the observations (Hobbs et al., 1993). The difference from deduction is that in abduction we "use" the rules of inference in the reverse direction. Namely, to use again Figure 1, we know the effects and the rules and we are looking to find the causes. We can either be looking for all possible causes or the most likely causes or the causes that (for some definition) optimally "explain" the effects.

[1]The solution is yes.

Abductive reasoning is most intuitively explained as the process that the fictional character Sherlock Holmes employs when solving a mystery. First he finds a collection of possible events that has as a consequence all observed facts. Then, if several such collections are possible he makes tests to eliminate all but one possibility - “when you have eliminated the impossible, whatever remains, however improbable, must be the truth”. We do abductive reasoning when we solve a Sudoku puzzle - we are given the puzzle and the constraints that the numbers in the puzzle satisfy and we are trying to recover the missing numbers. Of course, in a Sudoku puzzle, someone has made sure that there is only one possible solution. Life is rarely so obliging.

Inductive reasoning, induction or inductive inference is the process when we are given a set of observations of which some are causes and others effects and we would like to find rules that match the causes with their effects. Using Figure 1, inductive reasoning is when we are not given the rules and we try to “recover” them. Furthermore, we look for rules, or inferences, that generalise beyond the observations that have been used to generate them. We use inductive reasoning when we are trying to solve the puzzle given on Figure 2. We need to find the function f(xi , yi) and solve 8 + 11.

$$
\begin{aligned}
x_i + y_i = &\quad f(x_i, y_i) \\
1 + 4 = &\quad 5 \\
2 + 5 = &\quad 12 \\
3 + 6 = &\quad 21 \\
8 + 11 = &\quad ?
\end{aligned}
$$

Figure 2: An inference puzzle. File: puzzle.png

Automated reasoning was one of the first subdisciplines tackled in AI (Robinson and Voronkov, 2001). Deductive and abductive automated reasoning was applied very successfully in the development of expert systems which saw their rise in the 1980-1987 period. Expert systems are programs that solve complex problems by reasoning through information represented mainly as if–then rules (Jackson, 1998). Programs that use some knowledge and logic reasoning to solve complex problems are typically called knowledge-based systems.

The big breakthrough, particularly in the commercial sense, of inductive reasoning was accomplished in the AI subdiscipline of machine learning, particularly in the area of supervised machine learning. Machine learning is concerned with the problem of improving the

performance (of a program) on given tasks with respect to a performance measure, by using observations about the world (Mitchell, 1997).

Supervised machine learning simulates inductive reasoning. In supervised machine learning the task on which performance is improved is specified as follows. Given is a training set (a set of examples) of input output pairs (X1, y1),(X2, y2), . . . ,(XN , yN), where each of the yi are assumed to be generated by some unknown function y = f(x). The task is to discover a function h, called a hypothesis, that approximates the true function f. Why discover h and not f? Sometimes the examples in the training set are the observations from a stochastic process and the function f does not really exist outside of a particular set of instances. It could be that we are not certain that the elements of Xi that are included in the pairs are sufficient to uniquely define f, as is the case in the example on Figure 2.

For completeness we need to briefly introduce the other two types of machine learning: unsupervised learning and reinforcement learning. Supervised learning is so called because a human evaluator can determine whether the machine assigned label of a data point is correct. Unsupervised learning refers to a collection of methods which aim to draw inferences from data sets of unlabeled data or data that cannot be labeled. Unsupervised learning is typically used to find new patterns in data rather than learn how to use an existing pattern, which is what supervised learning does (Celebi and Aydin, 2016).

Reinforcement learning, introduced by (Barto et al., 1983), is perhaps closest to what we intuitively consider learning. It is concerned with the problem of searching for an optimal list of choices that need to be taken to accomplish a goal. This goal pursuit is represented as a function, so called reward function, whose value is to be maximized. The outcome of the choices taken is not known in advance and their payoffs are discovered by making them. The number of trials that need to be taken can be considerably large. One challenge in reinforcement learning is to correctly specify the reward function so that the right goal is achieved. Also, one has to balance out the exploring of new combinations of actions with the exploration of learnt payoffs (Sutton and Barto, 2018).

3 Subdisciplines of AI

At this point it is good to see how automating reasoning fits within that definition by Poole and Mackworth (2017) of synthesising artificial agents that behave intelligently. People do not only reason, they also determine what causes, effects and rules they choose to reason with (see Figure 3).

Figure 3: A model of automated reasoning and other cognitive tasks that support automated reasoning. File: aareason.png

A human does cognition by reasoning but also by doing all of the tasks that support reasoning which involve processing of input from their environment, actively looking or disregarding information, interacting with other people, sharing information, acting upon information etc. When a person plays chess for example, they look at the chess board and its state, consider

how this state can change, consider rules and strategies, consider the play of their opponent, plan moves and physically move pieces on the board. A person can also narrate what they are doing and give explanations and justifications for the choices they have made. In AI each of these cognitive tasks is studied by separate sub-fields. An AI method is typically developed to do one specific task, such as search for a good chess move, etc. When that method is applied, people step in to execute the related cognitive tasks that support it, such as represent the visually perceived state of the board into machine processable data.

In both automated reasoning and in machine learning, for the automation of reasoning to be possible, the information we want to reason about has to be represented in a way in which a computer program will be able to use. The subdiscipline focused on researching ways to do this is knowledge representation. In automated reasoning we typically use atomic representation using some type of formal logic. In machine learning, factored representation is used: a phenomenon is represented as a set of features (or rather a vector of feature values). Thus in an example pair (Xi , yi) the Xi is a set of features that represents a phenomenon. For example, if we are trying to learn how to distinguish apples from plums, we can describe each fruit with the features: size, colour, number of stones, softness. The value yi, called label or class, will be one of “apple” or “plum”. The hypothesis function that is generated is sometimes called a prediction model. The software that implements the prediction model is called a classifier.

Knowledge representation for the purpose of automated reasoning is done by professionals called knowledge engineers. Knowledge engineers collaborate closely with domain experts to identify and confirm the inference rules that are used by a knowledge based system. For the purposes of machine learning, the process of determining which features are relevant to be considered is called feature engineering. However, the process of describing examples in terms of feature values and particularly assigning the correct labels to phenomena described as sets of features can be done by anyone. For this task typically crowd computing is used. Simple cognitive tasks such as determining whether a fruit is an apple or plum, or whether a picture is of a face or not, that can be done by a person within seconds are called human intelligence tasks (HITs). The process of organising a large number of people to process a large number of HITs is called crowd computing. Online platforms offering crowd computing services have been developing to respond to the need for HITs processing, such as for example Amazon’s Mechanical Turk[2] and ClickWorker[3].

People transfer information among each other using a variety of natural languages. Natural language processing (NLP) is a subdiscipline of AI that studies information given in natural language, written or spoken, can be used by a computer program, and how information produced by a computer program can be communicated in a natural language. Examples of NLP include automated translation, speech-to-text conversion, question answering, and chatbots. Chat-bots are software applications that are able to engage in on-line chat conversation via text or text-to-speech. The topic of the conversation is limited to a particular

[2] https://www.mturk.com/
[3] https://www.clickworker.com/

topic. The responses are largely predetermined, matching to keywords from the input of the human participant, constructed from a structured base of information.

Vision is a sub-field that focuses on how video information can be processed by a machine for the purpose of knowledge representation and reasoning. Search and planning are subdisciplines of AI that study problem solving by artificial agents, specifically how an agent can generate and search through possible strategies and use them to build plans for accomplishing its goals. Lastly we have the AI sub-fields of robotics and multi-agent systems. Robotics does not need much introduction. This field is concerned with problems of artificial agents manipulating the real world and building both software and hardware for this purpose. Multi-agent systems studies the interactions of an agent with other agents in its environment.

It is important to observe that cognitive terms like “reasoning” and “learning” have different meanings when applied to a person and when applied to algorithms.

4. The threat of automating deduction and abduction reasoning

AI automates cognition by automating each cognitive task in separation and implements this automation by using people to support and execute the remaining related cognitive tasks. This approach imposes immediate limits to the abilities of AI. When a person does deduction or abduction, they can recognise when there is information missing. In the puzzle from the previous section that required us to find out if a married person is looking at an unmarried person there is not enough explicitly given information to confirm or refute that a married person is looking at an unmarried person. One needs to consider the assumption that a person is either married or unmarried but never both.

One of the main threats of using automated deduction is the problem of verifying that the persons specifying the causes and inference rules have “thought of and specified everything”. Otherwise, the deduction will not be correct. This is one aspect of a problem known in AI as the qualification problems: “The executability of an action can never be predicted with absolute certainty; unexpected circumstances, albeit unlikely, may at any time prevent the successful performance of an action.” (Thielscher, 2001).

The second main limitation of automated deduction (and abduction) is related to limits of what can be computed. For all deductions and abductions, automation is possible if the information is represented in a symbolic formal language. The number of computations we have to make depends exponentially on the level of detail we have used when the information is represented. Consequently, the “resolution” of representation is a kind of a “Goldilocks” problem - it has to be just right. However, even with the most careful curation of information, there are certain problems which simply cannot be processed with our current approaches to computation (Janota and Lynce, 2019).

The third main limitation of automated deduction and abductions is a financial one, closely related to the other two limitations, and frequently overlooked when discussing the AI applications of the past (Newquist, 1984). Representing information is a task that requires

extensive training of knowledge engineers and it has to be done in collaboration with domain experts who can determine what is important or not for a given problem. Both building a knowledge base and maintaining it requires a lot of expensive human hours from highly trained professionals.

5. The threat of machine learning

In machine learning, the problem of a huge requirement of expensive human labour is circumvented to an extent by the nature of inductive reasoning. Representing information for the use in machine learning is considerably human-easier to do than representation for deduction/abduction and can be done by anyone with little instruction. Most importantly, there is no need to identify inference rules, which is the most difficult and expensive component. This is not to say that there is no expertise required, however expertise here is not as necessary, which also is part of the risk that using machine learning involves. The problem of the artificial agent not being able to procure additional information remains. Feature engineers need to make sure all and every relevant feature is included.

When building a classifier we want a function that describes the training set, but we do not want it to “overfit” the prediction model to the training set. There is a great deal of uncertainty with respect to whether what we are using as a training set is an accurate and proportional representation of the world. For a lot of supervised machine learning problems, there is a “long tale” of feature value combinations that occur rarely. A classifier cannot be guaranteed to handle a so-called “black swan event”, namely correctly label a phenomenon that it has not encountered in the training set. It might seem advantageous to include as many features as possible when building classifiers, however, as was the case with automated deduction and abduction, there is a price to pay for a rich representation. A classifier that needs to handle a lot of features will be very slow to build, we say “training takes a long time”. The “Goldilocks” problem emerges again.

The most important thing to observe about inference rules discovered by machine learning is that these are rules of correlation not causation. Correlations can occur randomly and point to spurious relations between events[4]. When people do inductive reasoning they do not only observe that certain phenomena seem to be related to each-other, they further try to discern whether there is correlation or causation at play. This is done by: a) trying to find fringe examples that refute causation; b) by trying to generate an explanation for why the causation would hold; c) further confirm or refute that explanation empirically or analytically. Using inferences that are a product of correlation can be dangerous.

Nowhere is the threat of neglecting correlations as high as in using deep learning neural networks in the classification of images. Deep learning neural networks are perhaps the most famous approach to supervised learning. A simple neural network is given in Figure 4.

Figure 4: An example of a neural network. File: NeuralNet.eps

[4] For some interesting examples of spurious relations see https://tylervigen.com/.

An artificial neuron[5] is a node in a neural network. It is implemented as a mathematical function, so called an activation function that calculates the output of that node given an input, or set of inputs. Activation functions output zero as long as the value calculated is smaller than some threshold value, and “fire” when that threshold is surpassed.

The feature values Xi are fed in the input layer of the neural network. In Figure 6 each example is represented with five features. Each possible label is assigned its own neuron in the output layer. In Figure 6 a single label y is assigned or not, based on whether the output layer neuron “fires”. The layer in between the input and output layers is called a “hidden layer”. A neural network can have arbitrarily many hidden layers. The number of hidden layers and the number of neurons in each of the hidden layers, is determined by machine learning engineers in response to the specific machine learning problems that are being tackled.

The “training” of a neural network is the process that builds the prediction model. It consists of identifying the values for each of the “weights” of vertices that connect the neurons. The higher the weight the more relevant that particular input. The “training” of a neural network is mathematically the process of finding a local minimum of a function and it is done with a procedure called back-propagation (Rumelhart et al., 1986). In Figure 4 there are 54 weights that need to be “learned” and this is an unrealistically small network. The amount of training data needed increases with the size of the network.

The term “deep learning” means learning by neural networks that use hidden layers. Intuitively, the purpose of a hidden layer of a neural network is to allow not only for the feature values to be considered in the prediction model, but also their arbitrary combinations. However, using hidden layers also “breaks down” the possibility for intuitive interpretation of what each input and output of a neuron corresponds to. This lack of possibility for intuitive interpretation is what is referred to when neural networks are called “black boxes”.

This lack of intuitive interpretability of how the value of each feature impacts the classification is further exacerbated when the classification subject is an image. A special type of neural network architecture is used for image classification called convolutional neural networks (LeCun and Bengio, 1998). The input to a convolutional neural network are not engineered features but the pixel brightness values of the image.

The way a neural network “identifies” an entity on an image is by looking for correlations among pixels and not the way we as people do it. A person would recognise a cat because it has fur and pointy ears, or a specific colour and shape of tail.

The term “learning” can lead to potentially dangerous insinuations when we are talking about machine learning. When we say that a person learns, we intuitively consider the process of adapting our world view when exposed to new experiences and information. In supervised (and unsupervised) learning no such adaptation takes place. Once built, a prediction model does not

[5] Artificial neurons are also called “units”.

change with new information. In reinforcement learning, an artificial agent can continuously adapt the learnt set of actions until a particular optimum is achieved, but the reward function does not change once it is set by the human designer. And humans are not always good at precisely expressing what they want.

In summary, AI methods, like computer programming, are all solutions that require a machine to follow a rule. AI excels in the cognitive tasks in which the best approach is one that requires consistently and faithfully following a fixed, predetermined set of rules. We do not know how to program computers to break the rules, or even causally interpret them.

How to safely use AI knowing what it's limitations are? As a first quick rule-of-thumb, when replacing or enhancing a human role with an AI system, one can try to imagine if it is possible to outline the constraints of the role with rules and then what would it be like if the worker never makes any exceptions on those rules.

6 AI Ethics

Subfields in AI are being developed to directly respond to the issues arising from the use of AI. These are typically referred to collectively as AI ethics. There are at least seven disciplines in computer science that are clearly considered to be subdisciplines of AI Ethics. These are fairness (Mehrabi, 2019), algorithmic accountability Wieringa (2020), transparency (Diakopoulos, 2020), explainable AI (Gunning, 2019), privacy (Chapter 1, Kearns and Rroth, 2019)., responsible AI (Dignum, 2019), and machine ethics (Anderson and Anderson, 2011; Tolmeijer et al., 2020).

The AI ethics subfield represents a formidable effort towards safeguarding against the AI misuse. It suffers from two weaknesses that need to be overcome. The first weakness is best elaborated by Blackman (2020): AI Ethics often focuses on academic ideal solutions rather than implementable solutions. The second weakness is that focusing on preventing the bad does not directly bring about good. Namely, AI has a great potential to improve human lives without there being a direct financial benefit from this. We are researching only AI prevention but not AI empowerment.

7. Threats from the human-AI loop

To manage the ethical impact of an AI system, particularly in AI systems that involve some type of automated decision making, the need of a human supervisor of AI activities, or human-in-the-loop has been put forward[6]. Transparency and explainability efforts can be seen as an indirect way of maintaining such human oversight. The intuition is that if a person is in control then bad circumstances can be detected early and mitigated. However, the human-AI

[6] https://www.lawgazette.co.uk/law/ai-systems-must-have-a-human-in-the-loop-/ 5068587.article and https://finadium.com/ai-needs-human-in-the-loop-ethics-is-this-a-job-for-regulators/

system as a whole can be less safe than just human or just AI execution of tasks. Also, a human-AI system as a whole may end up performing worse than just people or just AI systems.

The human-in-the-loop concept comes from the field of supervisory control where it describes a system in which a human operator is a component of an automated control process (Sheridan, 2006; Rahwan, 2018). The human operator handles tasks such as supervision, handling exceptions, maintenance and others outside of the abilities of the automated process, but also serves as the locus of responsibility. The human control idea is central to the EU Ethics Guidelines for Trustworthy AI[7] that emphasizes the need for keeping users informed and in control.

There are numerous risks that arise from using human-in-the-loop and humanin-control as it has been long known in automation (Vincenzi et al.,2005). A person whose attention is not constantly on a task, which is likely to happen during supervising an automated activity, cannot be efficiently alerted to take control. Machine reaction times are a lot faster than that of a human. Consequently, a person cannot continuously follow what an artificial agent is doing. In turn, this disparity also makes it unfeasible that an agent always asks for human insight in ethically sensitive operations.

The human - AI dependency relationship when used within a deployed system can be corrupted from either side. On one hand human psychology can be exploited. For example, a user can be manipulated to consent to certain AI activities that they would not normally support (Forbrukerrådet, 2018). Regulation is ineffective when it requires that people's opinions are elicited without at the same time requiring that the elicitation is conducted without manipulative interfaces (Soe et al,2020).

On the other hand, a malicious actor can take advantage of the same human - AI relationship to take control of an AI system particularly when the system uses learning. This is particularly the case when reinforcement learning is used. An infamous example is the case of the Microsoft Tai chat-boot. This chat-bot was developed to respond to interaction on Twitter and learn new responses from the human input. It was very quickly "trained" to respond with inflammatory and offensive tweets (Hunt, 2016).

As people we can function with loosely defined roles of responsibility and oversight in our organisations. We are able to dynamically adapt to a situation and increase or decrease the control we have over the actions of others in a team, incentivize people to take on more responsibility or pay more attention. In a human-AI relationship, such loosely defined collaboration cannot work. Although the work envelope has been abandoned, a scaffolding is needed to take its place and define exactly which people with which minimal skills, under which circumstances are to provide input and feedback to an AI system that makes sensitive decisions.Clearly the biggest challenge remains for us to attain the wisdom of when to automate away a task completely, when to keep a person in control and when to leave it entirely on humans.

[7] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1

8 Concluding fears

Because AI aims to automate cognition, we tend to use a lot of anthropomorphic terms when we discuss it. The most considerable threat of AI is attributing to it human abilities that it does not have and overlooking the machine abilities that it does. The goal of this chapter has been to illuminate the differences between AI the science and AI the science fiction, by giving a gentle introduction on what AI is and how ethically sensitive issues are addressed within the field.

While one threat lies in attributing abilities to AI that it does not have, another lies in neglecting the abilities of simple programming code when executed in an open world. As Russel (2019) argues, it is not our ability to create sophisticated AI that is the biggest threat that we are facing, but our inability to test the effects on the world that a simple computer program has when this program interacts freely with other computer programs and people. For example, the algorithms that social media uses to determine whether the visibility of some content will be amplified or not are (likely) simple, but still shape how information is diffused around the globe in an unpredictable way.

No analysis of AI threats is complete without discussing superintelligence. A superintelligence is an artificial agent that is able to cognitively do all that a person is able to do and in this aspect surpasses the best of humans (Bostrom, 2014). The discussions of superintelligence embody the deepest fears and the hardest threats: Are we on the path to create superintelligence? How can we make sure that its incentives are aligned with our own as humanity? As intriguing as these questions are, their answer is the domain of philosophy. From the viewpoint of computer science and engineering, the likelihood of developing a superintelligence calls for an unscientific speculation Kelly (2017). General intelligence is a discipline in AI concerned with the development of general artificial cognition, instead of automating separate cognitive tasks. It is not even clear that general AI is within our reach given the current approaches and achievements (Mitchell, 2019), however even the development of general AI does not imply that superintelligence is reachable.

As seductive as the idea of superintelligence in the future is, we run the risk of allowing it to distract us from the true critical aspects of AI applications and the changes they bring to our society today. We can now claim with certainty that AI applications constitute a disruptive technology. Every societal disruption has the potential to spiral out of control and change our lives in an undesirable way. Therefore it is important to assess the threats as well as the benefits of adopting the new technologies.

One very real concern is in the role that AI technology plays and will play in our labour markets. By design AI aims to automate away human tasks and entire jobs are likely to disappear as a result in the near future[8]. This will necessarily lead to hardship for some people, at least short term, and it is important that societies address this challenge. The measures that are missing

[8] You can quickly assess the likelihood that your job will be taken over by robots here https://willrobotstakemyjob.com/

are those of policy, politics and education: how to help and support people whose job disappeared because of automation?

The profile of the disappearing jobs is not difficult to outline given the skills of an artificial agent - repetitive jobs, hard to fill cheaply, that do not require the simultaneous use of several cognitive faculties. The rise of the AI also comes with new demands on people. Human labour is used every step of the way to make an AI system possible. This human aspect of the AI success is often not vigorously promoted (Zittrain, 2019). The new unskilled labour is poised to be the human cognitive labour force that does small human intelligence tasks for small income. Are our societies equipt to identify the rights of “small gig” workers and protect them from exploitation?

There is a real threat of over-hyping the abilities of AI applications. For AI researchers, this threat is well documented as the field has survived two “springs” of hype and two subsequent “winters” where disappointed stakeholders had withdrawn research funding (Moor, 2006a). Beyond the threat to research, over-hyping AI abilities by companies and thrill invoking writers can lead to a distorted public opinion on how much to trust an AI application, when and for what. There is only one sure way to mitigate this risk, and the risk of job insecurity - more AI education. This viewpoint is increasingly adopted by education systems around the globe.

Children using computers and smart devices with ease are not necessarily computationally literate children. Computer programming is increasingly being recognised as a subject to be included in primary school education (Serafini, 2011). The goal of such an education is not to turn everyone to a career of a computer programmer, but to enable everyone to use the tools of future professions. Following this trend is the argument that AI too should be introduced in primary school curricula (Chan, 2019).

AI, at present, is a subject in the university curricula of computational sciences. The voices of concern are louder for educating programmers not only to do AI but to consider the societal impact of AI. However these tend to be sometimes limited to the re-education of engineers, and do not include introducing AI to other “non-technical” curricula.

O’Neil in 2017 accused academia of being “asleep at the wheel, leaving the responsibility for this education to well-paid lobbyists and employees who’ve abandoned the academy”[9]. This has been shown to be rather incorrect, as universities too are making efforts to further educate their computational students in the ethical aspects of AI (Fiesler et al., 2020). The biggest concern in this educational efforts is not whether they exist but how efficient they are in developing this crucial skill of the future scientists and engineers to “not forget about ethical concerns when focused on technical systems issues"[10]. But we only have to teach the programmers how to be “humanists”, we also need to teach the “humanists” how to deal with computer programs.

[9] https://www.nytimes.com/2017/11/14/opinion/academia-tech-algorithms.html
[10] https://cacm.acm.org/magazines/2019/8/238345-embedded-ethics/fulltext

For social scientists, educationists, and psychologists there are plenty of questions opened by introducing AI systems in our everyday lives. We are only now waking up to the reality that we build technology to meet our needs, but we, how we live and what we value, are also changed by this technology in turn as well (Rettberg, 2014). AI helps BigData transform data into knowledge. This means that all those surveillance camera videos can now be analysed in real time, what we do online can be known to persons unknown, our whereabouts, our habits, can all be made visible. A lot of our identity is related to how we are seen by our peers, how much we are seen by our peers. However, if everyone is visible all the time, there might be a fear of not being seen at all that can be a new incentive that changes who we are as individuals (Bucher, 2012). To study these phenomena, humanists and social scientists also need to learn at university what AI is and how it works “under the hood”.

If AI is an existential threat to humanity, then the threat is more likely to come in the shape of thousands of paper cuts that we fail to address rather than that of Terminator’s Skynet. We discussed some societal fears associated with the role of AI in our society today and in our immediate future. However, the threat of AI is not one that can be met by a task force and the diligence of few academic guardians, the same way that fire damage is not prevented by establishing a fire department. Safe AI use and development, AI that empowers individuals and societies, is the product of continuous education and individual responsible choices.

References

Algorithmic Watch (2020). Ai ethics guidelines global inventory. https://inventory.algorithmwatch.org/. [Online; accessed 20-May-2020].

Anderson, M. and Anderson, S. L. (2011) eds., Machine Ethics, Cambridge University Press.

Bellman, R. E. (1978). An Introduction to Artificial Intelligence: Can Computers Think? Boyd & Fraser Publishing Company.

Black, R. (2020). A Practical Guide to Building Ethical AI. Harvard Business Review. https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Bucher, T. (2012). Want to be on the top? algorithmic power and the threat of invisibility on facebook. New Media & Society, 14(7):1164–1180.

Celebi, M. E. and Aydin, K. (2016). Unsupervised Learning Algorithms. Springer Publishing Company, Incorporated, 1st edition.

Chan, D. (2019). Primary students to be taught AI in Guangzhou schools government has decided to give priority to the development of AI, information and biopharmaceutical industries.

https://asiatimes.com/2019/07/primary-students-to-be-taught-ai-in-guangzhou-schools/. [Online; accessed 10-June-2020].

de Kleijn, M. (2018). Using AI to map . . . AI? Elsevier Connect.

Diakopoulos, N (2020). The Oxford Handbook of Ethics of AI, M. D. Dubber, F. Pasquale, S. Das, eds. Oxford University Press,

Dignum, V. (2019). Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer.

Fiesler, C., Garrett, N., and Beard, N. (2020). What do we teach when we teach tech ethics? A syllabi analysis. In Proceedings of the 51st ACM Technical Symposium on Computer Science Education, SIGCSE '20, page 289–295, New York, NY, USA. Association for Computing Machinery.

Gunning, D. and Aha, D. (2019). Darpa's explainable artificial intelligence (xai) program. AI Magazine, 40(2):44–58.

Hobbs, J. R., Stickel, M., Martin, P., and Edwards, D. (1993). Interpretation as abduction. Artificial Intelligence Journal.

Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. https://goo.gl/mE8p3J. [Online; accessed 10-June-2020].

Jackson, P. (1998). Introduction to Expert Systems. Addison-Wesley Longman Publishing Co., Inc., USA, 3rd edition.

Janota, M. and Lynce, I., editors (2019). Theory and Applications of Satisfiability Testing - SAT 2019 - 22nd International Conference, SAT 2019, Lisbon, Portugal, July 9-12, 2019, Proceedings, volume 11628 of Lecture Notes in Computer Science. Springer.

Kearns, M. and Roth, A. (2019). The Ethical Algorithm: The Science of Socially Aware Algorithm Design. Oxford University Press.

Kelly, K. (2017). The Myth of a Superhuman AI. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/. [Online; accessed 10-June2020].

Kleiner, K. (2009). Why smart people do stupid things intelligence by itself doesn't make you rational. Thinking rationally demands mental skills that some of us don't have and many of us don't use. University of Toronto Magazine.

LeCun, Y. and Bengio, Y. (1998). Convolutional Networks for Images, Speech, and Time Series, page 255–258. MIT Press, Cambridge, MA, USA.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning.

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux.

Mitchell, T. (1997). Machine Learning. McGraw Hill.

Moor, J. (2006a). The Dartmouth College Artificial Intelligence Conference: The Next Fifty Years. AI Magazine, 27(4):87.

Newquist, H. (1984). The Brain Makers : Genius, Ego, and Greed in the Quest for Machines That Think. Sams Publishing.

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

O'Neil, C. (2016). Weapons of Math Destruction. Crown Books.

Poole, D. and Mackworth, A. (2017). Artificial Intelligence: Foundations of Computational Agents. Cambridge University Press, Cambridge, UK, 2 edition.
Popejoy, A. B., Ritter, D. I., Crooks, K., Currey, E., Fullerton, S. M., Hindorff, L. A., Koenig, B., Ramos, E. M., Sorokin, E. P., Wand, H., Wright, M. W., Zou, J., Gignoux, C. R., Bonham, V. L., Plon, S. E., Bustamante, C. D., and Clinical Genome Resource (ClinGen) Ancestry and Diversity Working Group (ADWG) (2018). The clinical imperative for inclusivity: Race, ethnicity, and ancestry (rea) in genomics. Human Mutation, 39(11):1713–1720.

Rahwan, I. (2018). Society-in-the-loop: programming the algorithmic social contract. Ethics and Information Technology, 20(1):5–14.

Rettberg, J. W. (2014). Seeing Ourselves Through Technology: How We Use Selfies, Blogs and Wearable Devices to See and Shape Ourselves. Palgrave.

Robinson, A. and Voronkov, A., editors (2001). Handbook of Automated Reasoning. Elsevier Science Publishers B. V., NLD.

Russell, S. (2019) Human Compatible, AI and the Problem of Control. Penguin Random House UK.

Russell, S. and Norvig, P. (2015). Artificial Intelligence: A Modern Approach. Pearson Education, 3 edition.

Serafini, G. (2011). Teaching programming at primary schools: Visions, experiences, and long-term research prospects. In Kalas, I. and Mittermeir, R. T., editors, Informatics in Schools.

Contributing to 21st Century Education, pages 143–154, Berlin, Heidelberg. Springer Berlin Heidelberg

Sheridan, T. B. (2006). Supervisory Control, chapter 38, pages 1025–1052. John Wiley & Sons, Ltd.

Soe, T. H., Nordberg, O. E., Guribye, F., and Slavkovik, M. (2020). Circum-vention by design - dark patterns in cookie consent for online news outlets. In Lamas, D., Sarapuu, H., Larusdóttir, M., Stage, J., and Ardito, C., editors, NordiCHI '20: Shaping Experiences, Shaping Society, Proceedings of the 11th Nordic Conference on Human-Computer Interaction, Tallinn, Estonia, 25-29 October, 2020, pages 19:1–19:12. ACM.

Sutton, R. S. and Barto, A. G. (2018). Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–9.

Thielscher, M. (2001). The Qualification Problem: A Solution to the Problem of Anomalous Models. Artificial Intelligence, 131(1):1 – 37.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., and Bernstein, A. (2020). Implementations in machine ethics: A survey. CoRR, abs/2001.07573.

Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 1–18, New York, NY, USA. Association for Computing Machinery.

Zhang, B. and Dafoe, A. (2020). U.S. public opinion on the governance of artificial intelligence. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20, page 187–193, New York, NY, USA. Association for Computing Machinery.

Zittrain, J. (2019). The hidden costs of automated thinking. The New Yorker. https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking.