ARTIFICIAL INTELLIGENCE

Fredrik Heintz

Introduction

Artificial intelligence (AI) is a general-purpose technology that will impact most, if not all, aspects of both our society and our personal everyday life. AI technology has enabled applications such as speech interfaces, vision-based object recognition, and machine translation. AI technology also makes recommendations about music, books, and movies for you, decides whether you will get a bank loan, and controls what posts you see on social media, all of which can have a major impact on your life. It is clear that AI technology will play a central role in most aspects of our professional and private lives as well as society at large. Kevin Kelly predicted that “The business plans of the next 10,000 start-ups are easy to forecast: Take X and add AI” (Kelly 2016). Andrew Ng says that AI is the new electricity, it is a fundamental part of almost everything (Lynch 2017).

The field of artificial intelligence was founded in 1956 with the famous Dartmouth conference. From the start, the goal was to make computers do everything that humans can do. To quote the proposal:

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

(McCarthy et al. 1955)

However, the quest for understanding and mechanising human intelligence started much earlier (McCorduck 2004; Nilsson 2010). For example, Aristotle studied what reasoning is and came up with three types of reasoning – deductive, abductive, and inductive – which covers much of what is being done today in both machine reasoning and machine learning.

From a pragmatic perspective, an AI system can be described as a system that receives data, makes decisions, and acts on this information. Sense-Plan-Act is another way to describe such
a system. It senses the environment, it makes a plan for how to achieve its goals, and then acts based on the plan (Russell and Norvig 2016). In many cases, these systems can learn from data to improve over time. They are often called agents, as they have a sense of agency which differentiates them from other computer programs. This also gives rise to a cognitive and social view on computation (Shoham 1997).

Machine learning is currently seen as the most interesting part of AI, both because many consider it an essential part of intelligence and because it allows computer programs to improve over time based on experience. This is important since it is hard for people to specify exactly what we want a system to do. Instead, the machine can partly learn what to do and how to do it, as well as improve over time, by collecting data and modifying its behaviour (Brynjolfsson and Mitchell 2017).

An interesting example is AlphaGo which plays the Chinese board game Go better than any other human. In 2015, it beat one of the best humans through a combination of learning from existing games and self-play (Silver et al. 2016). In 2017, an improved version called AlphaGo-Zero beat its previous version in 100 games out of 100 purely through self-play (Silver et al. 2017). The same system can also play chess better than all existing chess playing programs. The same type of techniques has also been used by Google to reduce the energy consumption in their data centres by 40%.

By observing the human expert commentators, and most importantly, their surprise at some of the moves, it is clear that AlphaGo does not purely replicate human strategies, rather it extends existing strategies and creates new ones. This leads to interesting questions about how to validate decisions made by an AI system and how to maintain meaningful human control over a system capable of making better decisions than we are.

The purpose of this chapter is to introduce artificial intelligence (AI) from a non-technical perspective. After reading this chapter you should have a basic understanding of what AI is and where it comes from; what the main concepts and methods are; some of the applications; and some of the future research challenges.

What is AI?

Artificial intelligence is about understanding intelligence sufficiently well to be able to recreate it in machines. Another way to describe AI is as systems taking input, analysing this data, making decisions, and then acting based on these decisions. This approach is often called the Sense-Plan-Act approach (Russell and Norvig 2016). In many cases, these systems learn to improve their performance over time either from data given to them or from data collected during its execution.

A challenge with the definition of AI is that there is no commonly agreed upon definition of human intelligence (Legg and Hutter 2007). A computer can often do things that we assume requires intelligence without any effort, like solving difficult mathematical problems. At the same time, computers are often very poor when it comes to doing what appears to be really simple things like learning a new concept from abstract descriptions, like a zebra is a horse with black-and-white-striped fur, or manipulation tasks, like folding a blanket or tying a pair of shoes. A consequence is the recognition that human intelligence is not necessarily the right baseline to compare against. There is an inherent anthropocentrism in the word artificial, assuming that intelligence can only originate in a human body.

1 https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40 [10.03.2021]
Therefore, the focus is often on building systems that behave intelligently, rather than claiming that the systems are intelligent. It is also common to focus on particular cognitive functionalities such as planning, natural language processing, or perception. Cognitive science, on the other hand, tries to understand how human cognition works and also tries to replicate the functionality through computer models, which then aim at emulating human cognition with both its strengths and weaknesses (Thagard 2008). The engineering approach to AI, on the other hand, is to develop methods, algorithms, and programs that exhibit intelligent behaviour based on computer science rather than cognitive science or neuroscience, even though nature is often an important source of inspiration. Classical examples of this are artificial neural networks and genetic algorithms. Artificial neural networks were developed in the 1950s based on a simple model of how researchers thought that human neurons worked (Hubel and Wiesel 1962; Rumelhart et al. 1986) and are today implemented as high dimensional matrix operations running on massively parallel GPUs.

AI has many subfields that study different aspects of intelligent behaviour and cognition. Common topics at the main AI conferences include machine learning, knowledge representation and reasoning, heuristic search, planning and scheduling, natural language processing, computer vision, robotics, and multi-agent systems. All of these topics have been studied since the 1950s. Most of them were in fact discussed already at the seminal Dartmouth conference in 1956.

Two of the most important subfields are machine learning and knowledge representation and reasoning. Knowledge representation and reasoning is the scientific study of how to represent knowledge in a computer and how to reason with this knowledge to draw valid conclusions. Machine learning is the scientific study of how a computer can learn things such as finding patterns, recognising objects, and acting to achieve specific goals. Machine learning is mostly based on statistics and correlations (black box models) while knowledge representation and reasoning are mostly based on explicitly modelling cause and effect (white box models). One of the major consequences is that machine learning is mostly data driven while reasoning is mostly knowledge driven, where explicit knowledge has to be elicited from domain experts. The second major consequence is related to interpretability, explainability, and thus trustworthiness and reliability, where white box models by design provide an explicit representation of the reasoning that facilitates understanding, while black box models are very hard to understand, as the rules are implicit and often distributed in the model.

Currently, most of the attention is focused on machine learning, while knowledge representation and reasoning were the focus in the 1980s and 1990s, often in the form of expert systems. The next big step is likely the combination and integration of reasoning and learning, maybe in a similar manner as we humans do it with two separate but somehow connected systems (Kahneman 2011). System I is the fast, automatic, and opaque system for perception and intuition with very limited introspection, which shares many similarities with data-driven machine learning approaches. System II is the slow, deliberate, and explicit system for analytical thinking and planning with a high degree of introspection, which corresponds roughly to formal, symbolic, reasoning-based approaches. Another significant trend is to study the implications of AI and to make sure that AI is developed in a way that benefits all. The EU is for example putting its weight behind the concept of Trustworthy AI, which requires AI systems to follow the applicable rules and regulations, live up to four ethical principles, and to have a robust and safe implementation (High-Level Expert Group on AI 2019). A consequence of this is that the field of AI is broadening and today includes researchers from a wide variety of scientific disciplines, not only computer scientists.

The rest of this chapter will provide more details about the main topics of reasoning, learning, Trustworthy AI, and human–AI interaction.
Reasoning

Reasoning is mainly about inferring implicit information. An everyday example is to fill in the missing numbers in a game of Sudoku. Another, classical, example is the logical rule of *modus ponens*, which, given two facts — (1) $p$ implies $q$ and (2) $p$ — allows the conclusion that $q$ holds to be drawn. It is not explicitly stated, but it is implicitly true given the standard interpretation of logic. The fact that $q$ holds is called a *consequence* and that the conclusion is *entailed* by the two statements. This is an example of classical logic. The study of logic is basically the study of what valid conclusions can be drawn from a set of statements and its roots go back to the ancient Greeks.

The study of formal logics has also provided us with many new insights and scientific facts of formal systems such as G"odel's incompleteness theorem that states that any formal system that is powerful enough to encode the natural numbers will contain truths that cannot be proven by the system (Arbib 1987). This means that there are known truths that cannot be proven using the system; in other words, it is *incomplete*. Another important result is that first-order logic is *undecidable*, that is, it is not generally possible to prove that a given first-order statement is true or false relative to a set of first-order statements.

Reasoning is often related to knowledge representation. The question is basically: what is knowledge, how can we represent knowledge in a computer, and what can we do with this knowledge? (Brachman and Levesque 2004).

Five important types of reasoners are SAT-solving that encode problems as satisfiability problems (Biere et al. 2009), CSP-solving that encode problems using finite domain constraints (Fruhwirth and Abdennadher 2003), model checking that check if a system often described as a timed automaton satisfies temporal logical properties such as safety and liveness properties (Clarke et al. 2018), automated theorem proving that prove logical statements (Newborn 2000), and planning that find sequences of actions that satisfy a goal given a domain model (Ghallab et al. 2004).

The most common technique to solve reasoning problems is *search*. It is a systematic way of trying different combinations until a solution is found. With a complete search, every possible combination is eventually tried. A concrete example is filling in an empty cell in a partially filled Sudoku. Try 1, check if the row, column, and square conditions are satisfied. If they are, then fill in the next cell, otherwise, try the next number. If there are no more numbers to try, then the partial Sudoku is not correctly filled in and cannot be solved. Through a technique called *backtracking*, a general procedure for solving Sudokus can be constructed which fills in the empty cells one by one until one of them cannot be filled in, in which case the algorithm has to backtrack by “unfilling” the previous cell and restarting the search from there. Another common technique is *mini-max* search for solving two-player games.

Formal reasoning is very useful and powerful in those applications where the domain can be represented formally using logics, finite domain constraints, or similar. Examples include formal verification of microprocessors, communication protocols, planning elevator rides, and configuration management. The main research challenge is the trade-off between expressivity and efficiency – the more expressive the formal language, the longer it takes to compute the answers, and in many cases, it is not even possible to guarantee an answer. Due to this, an important scientific endeavour is to find what is called *maximally tractable subsets of logic*, which allow all valid conclusions to be drawn for an infinite time for some fragment of the original logic (and that the fragment is as large as possible). From an application point of view, the main challenge is usually how to formally model the thing of interest.
Learning

Learning is fundamentally about how to improve given information.

Machine learning can be divided into three areas: supervised learning, unsupervised learning, and reinforcement learning (Bishop 2007). Supervised learning techniques are by far the most used in practice. Based on large collections of input-output examples, called training data, these methods try to find an underlying model that generalises the set of training data until it achieves acceptable performance on a different set of test data. To verify that the model actually generalises beyond the training and test data, its final performance is evaluated against validation data.

The three most common model types are classification, regression, and generative models. A classifier is a model that takes an individual, usually described by a vector of features, a feature vector, and determines which of a finite set of classes it belongs to, or, classifies it. Object recognition is a classic example of classification. A regression model is a mapping from an input domain to an output domain, which means that it is capable of computing the output for every possible input. Estimating the price of a good based on a feature vector or predicting the weather tomorrow based on the weather the last seven days are two common examples of regression models. Common regression models are neural networks, support vector machines, and Gaussian Processes. The third type, generative model, learns a probability distribution over the features, which means that it is possible to generate instances from the model, hence the name generative. The two most common generative models are Variational Auto Encoders (VAE) and Generative Adversarial Networks (GAN) (Goodfellow et al. 2014).

Deep learning is probably the technique that has received the most attention. Normally, deep learning refers to an artificial neural network with many layers (Goodfellow et al. 2016). A neural network consists of two or more layers of neurons connected by weights. The network is trained by changing the weights between the neurons. In a feed forward neural network, the neurons in layer \( n \) are connected to neurons in layer \( n + 1 \). In a recurrent neural network (RNN), neurons can be connected both to the next layer and to themselves. An important type of recurrent neural network is the Long Short-Term Memory (LSTM) which is very good for learning sequences to sequence models, where one sequence is mapped to another sequence such as translating between languages (Hochreiter and Schmidhuber 1997).

Neural networks have been studied since the 1950s (Schmidhuber 2015), starting with the Perceptron. However, at this time, it was not known how to train neural networks with more than two layers. As it was shown in 1969 that these two layer networks, the Perceptrons, could not represent non-linear functions such as XOR (Minsky and Papert 1969), the interest in neural networks dramatically decreased until the mid-1980s when backpropagation for propagating the error backwards through several layers was developed (Rumelhart et al. 1986). Backpropagation was actually first discovered by a Finnish master’s student (Linnainmaa 1970). In the early 1990s convolutional neural networks (CNN) was developed to effectively represent dynamic filters often applied in computer vision (LeCun et al. 1990). Neural networks have been used commercially, for example for character recognition since the 1990s, but it took another 20 years before its big breakthrough. The reason is that a large number of training examples is needed which requires significant computational resources to train. The major breakthrough came with AlexNet (Krizhevsky et al. 2012) which significantly outperformed all existing object recognition methods on the ImageNet challenge (Deng et al. 2009) after several successive improvements (LeCun et al. 2015).

Deep neural networks have had a profound impact on many areas including computer vision, speech recognition, machine translation, and natural language processing (LeCun et al. 2015; Brynjolfsson and Mitchell 2017).
Reinforcement learning is a technique for learning how to act in an environment through trial and error. The basic idea is that the system learns a mapping from state action pairs to expected cumulative rewards, representing an estimation of how good it is to perform a certain action in a certain state. To update the mapping, the system interacts with the environment by trying different actions and receiving positive or negative rewards. Learning to ride a bike is a good example of reinforcement learning, where you learn how to control your body to not fall while riding. Since the size of the state-action mapping in most applications is too large to be explicitly represented, it is often approximated used a deep neural network, called deep reinforcement learning. Reinforcement learning is for example used in robotics but also to play computer games such as Atari (Mnih et al. 2015), Go (Silver et al. 2016), and Star-Craft (Vinyals et al. 2019).

Even though machine learning techniques such as deep learning have been very successful in tasks such as classifying objects, playing games, and translating between languages, there are also many challenges. Some of them are to learn more general representations (Bengio et al. 2013), to avoid overfitting (Srivastava et al. 2014), to learn multiple things at the same time (Caruana 1997), to explain why the network made certain recommendations (Gunning et al. 2019), to combine learning and reasoning (Bottou 2014), and to combine learning and logics (Richardson and Domingos 2006). One open question is whether it is sufficient to use neural networks, or if some innate knowledge, structures, or algorithms are also required, to learn everything that we humans can do (Marcus 2018; Dietterich 2017).

Trustworthy AI

To maximise the opportunities and minimise the risks, Europe has decided to focus on human-centred Trustworthy AI based on strong collaboration among key stakeholders. Trustworthiness is a prerequisite for people and societies to develop, deploy, and use AI systems. Without AI systems – together with humans – being demonstrably worthy of trust, unwanted consequences may ensue and its uptake might be hindered, preventing the realisation of the potentially vast social and economic benefits brought by AI systems (High-Level Expert Group on AI 2019).

Trust in the development, deployment, and use of AI systems concerns not only the technology’s inherent properties but also the qualities of the socio-technical systems involving AI applications. Analogous to questions of (loss of) trust in aviation, nuclear power, or food safety, it is not simply components of the AI system but the system in its overall context that may or may not engender trust. Striving towards Trustworthy AI hence concerns not only the trustworthiness of the AI system itself, but requires a holistic and systemic approach, encompassing the trustworthiness of all actors, and processes that are part of the system’s socio-technical context throughout its entire life cycle (High-Level Expert Group on AI 2019).

According to the High-Level Expert Group on AI, Trustworthy AI has three main aspects, which should be met throughout the system’s entire life cycle (High-Level Expert Group on AI 2019):

1 It should be Lawful, ensuring respect for all applicable laws and regulations;
2 It should be Ethical, ensuring adherence to ethical principles and values; and
3 It should be Robust, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Each of these three components is necessary but not sufficient on its own to achieve Trustworthy AI. Ideally, all three reinforce each other. However, in practice, there may be tensions between these elements, for example when breaking the law might be necessary to save lives or...
when the scope and content of existing law might be misaligned with ethical norms. According to the ethical guidelines, it is our individual and collective responsibility as a society to work towards ensuring that all three components help to achieve Trustworthy AI. The ethical principles are *respect for human autonomy*, *prevention of harm*, *fairness*, and *explicability*. Based on the three components and the four ethical principles, the High-Level group defined seven key requirements for Trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination, and fairness, (6) environmental and societal well-being, and (7) accountability. To assist organisations using AI systems and developers to build Trustworthy AI systems, the High-Level group developed an assessment list for Trustworthy AI (High-Level Expert Group on AI 2020).

To achieve these requirements robustly there are many technical research challenges. Some of them are fairness, explainability, transparency, and safety.

*Fairness* can be defined as “absence of any prejudice or favouritism toward an individual or a group based on their inherent or acquired characteristics” in the context of decision-making (Mehrabi et al. 2019). AI-based complex socio-technical systems may amplify data biases, and also introduce new forms of biases (Osoba and Welser 2017). The reason is that AI systems usually rely on data, which may be biased in ways that are socially significant. One source of bias is that data generation is often a social phenomenon full of human biases. This bias may carry over to the decision-making of the AI system in ways that are unfair to the subjects of the decision-making process. For example, it has been shown that automated methods applied to language necessarily learn human biases inherent in our use of language (Caliskan et al. 2017). Fairness-aware machine learning algorithms seek to provide methods under which the predicted outcome of a classifier operating on data about people is fair or non-discriminatory. Broadly, fairness-aware machine learning algorithms have been categorised as those pre-processing techniques designed to modify the input data so that the outcome of any machine learning algorithm applied to that data will be fair, those algorithm modification techniques that modify an existing algorithm or create a new one that will be fair under any inputs, and those post-processing techniques that take the output of any model and modify that output to be fair. Many associated metrics for measuring fairness in algorithms have been explored (Mehrabi et al. 2019).

*Explainability* can be defined as “the act or process of making something clear or easy to understand”. A more specific definition by DARPA of Explainable AI is as “AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future” (Gunning et al. 2019). Explainable AI is essential if users are to understand, appropriately trust, and effectively manage AI systems. Post hoc and transparent design explanations can be used to help AI explain its strengths and weakness and convey how it understands a concept (Guidotti et al. 2018). In post hoc explanation, given an opaque AI model, a so-called *black box*, the aim is to reconstruct its logic either by mimicking the overall behaviour of the opaque model with a transparent classifier (*global explanation*) or by focusing on the construction of a local explanation for a specific instance (*local explanation*). On the other hand, transparent design explanations aim to develop a model that is explainable on its own. Some of the research challenges are developing methods for learning more explainable models, designing effective explanation interfaces, and understanding the psychological requirements for effective explanations (Miller et al. 2017). There are also challenges related to designing metrics to measure the performance of these methods.
Transparency in AI plays a very important role in the overall striving to develop more Trustworthy AI as applied to markets and in society. It is one of the seven key requirements for Trustworthy AI, and one of the key five principles emphasised in the vast number of ethical guidelines addressing AI on a global level (Jobin et al. 2019). It is particularly trust and issues of accountability that drive the contemporary value of the concept, including the narrower scope of transparency found in Explainable AI (Ribeiro et al. 2016; Miller 2019). AI transparency takes a system’s perspective rather than focusing on the individual algorithms or components used (Larsson and Heintz 2020). It is therefore a less ambiguously broad term than algorithmic transparency (Diakopoulos and Koliska 2017). In order to understand transparency in AI as an applied concept, it has to be understood in context, mitigated by literacies, information asymmetries, “model-close” explainability, as well as a set of competing interests. Transparency in AI, consequently, can best be seen as a balancing of interests and a governance challenge demanding multidisciplinary development to be adequately addressed (Larsson 2019; Larsson and Heintz 2020).

Safety. AI systems should be conceived to be safe for humans, be robust against perturbations, varying contexts, malicious attacks, and have fallback plans. As AI systems become more complex, to achieve safety and robustness we need to re-understand their evaluation to (1) verify a system under acceptable assumptions (verifiability), or (2) precisely assess how often, how much, and when the system may fail (calibration, profiling, and context-dependent evaluation) (Hicks 2018). One very important area is the area of safe reinforcement learning (García and Fernández 2015) where an agent or robot learns to achieve goals in a guaranteed safe manner. In areas where verification is not possible or practicable, AI can be used to systematically validate other AI systems through simulating the interaction of AI systems with realistic, AI-learned models of the environment that an AI system will have to interact within reality (Dahmen et al. 2019).

A promising direction of active research is to develop methods for learning causal models which allows us through reasoning to explain the past and to predict the future (Pearl and Mackenzie 2018). It is also crucial that these causal models can be reasoned about in order to analyse, for example, their fairness properties. For this to be achieved, the major scientific challenge is how to integrate learning and reasoning in a principled manner while retaining the explanatory power of more structured, often logical, approaches together with the adaptability and efficiency of data-driven machine learning approaches. This is also in line with how people think, using two different systems (Kahneman 2011). Many data-driven machine learning approaches could be seen as example instances of System I, which often work but without an explicit understanding of exactly what they base their decisions on. On the other hand, most knowledge-based machine reasoning approaches could be seen as instances of System II, which often produce high-quality explicit and formally guaranteed results, but only if the encoded knowledge is correct and representative of reality. A major research challenge is how to combine these different methods into working AI systems.

Human-AI interaction

An interesting question related to AI is how this influences the role of humans. Humans and computers are fundamentally good at different things, which makes humans and computers complementary (Kamar 2016). Instead of complete automation, where we hand over the control completely to the computer, it is better if humans and computers solve problems together. Even if a computer is good at recognising objects and classifying images, humans are many times still even better and definitely more general. The role of humans then becomes to train and teach AI algorithms right from wrong and monitor that they are actually doing the right thing. The training most likely never will be completely finished, but rather incremental and
Artificial intelligence

continuous as new concrete examples of incorrect decisions and situations where the computer does now know what to do are collected. In these cases, we humans have to take over and provide the correct answer. A challenge for us humans then becomes what we think is right, given our different perspectives and backgrounds.

An interesting example is chess. We humans have no chance against the best chess computers, and have not had a chance for over 20 years (Siegel 2016). At the same time, the quality of human chess playing is increasing, as we are practising against chess computers. Some claim that Magnus Carlsen is the best chess player in the world since he is the human who is the best at playing like a computer. This is natural to him as he has been practising against the computer since he was a young child. What is even more interesting is that if you combine humans and chess computers, the joint team, called a centaur, becomes better than both the best humans and the best computers. It is even the case that the team becomes even better if you include several people (Kasparov 2017). This is a concrete example of how the result improves when humans and computers collaborate to solve complex problems. There is no dichotomy between humans and computers, it is not a question of either/or, but rather humans and computers. Simplified, one can say that computers are good at doing, while humans are good at what should be done and why. We are good at asking questions, and computers are good at answering them. Examples are question answering systems that are great at answering questions and planning systems that can generate elaborately detailed plans for how to achieve goals, but the questions and the goals have to be provided to the systems by human users.

An important observation is that it is a different skill to play chess with a computer compared to playing chess on your own. This means that even if you are a really good expert and you are provided with the best possible tool, it is not necessary that the result improves. You might still perform worse compared to a person who is less of an expert in the subject but more of an expert on using the tool effectively. To really leverage the computational power, we need to both educate people in solving problems with AI tools and adapt the way we work to truly leverage the tools. Thus, relevant education, changed ways of working, and new organisational forms are required; see for example Wagner (2020a). A central capability is to transform business problems into computational problems. That is, to formulate problems in such a way that computers or computer tools can assist (Brynjolfsson and Mitchell 2017).

The social and economic pattern that derives from the complementary nature of human and machine intelligence is one that has been prevailing and gaining momentum since the beginning of the industrial revolution: specialisation and division of labour. Specialised machine intelligence allows this process to further accelerate, especially in the field of knowledge work, leading to micro-division of labour (Wagner 2020b).

By interpreting AI as a new type of agent within a company, it has been argued that a number of effects on companies can be supported (Wagner 2020c):

1. AI intensifies the effects of economic rationality on the firm;
2. AI introduces a new type of information asymmetry;
3. AI can perforate the boundaries of the firm;
4. AI can create triangular agency relationships; and
5. AI has the potential to remove traditional limits of integration.

Computational thinking captures this general skill of solving problems in a way that computers can assist (Wing 2006, 2011). For computers to help us, we have to be better at understanding how a computer solves problems. Thus, computational thinking is to a large extent about learning to understand how a computer “thinks” when it solves a problem.
When you solve problems with a computer, it is often about describing to the computer what should be done, rather than doing it yourself. Programs are descriptions of how to solve something that computers understand. Traditionally, humans have to describe every step of the process in great detail. AI actually reduces this by enabling the computer to fill in many of the details. It is also interesting to consider the macro-level. Instead of looking at the individual actor (micro-level), we take the perspective of groups rather than individuals. An interesting example is human-agent collectives which are presented as a new class of socio-technical systems in which humans and smart software (agents) engage in flexible relationships in order to achieve both their individual and collective goals (Jennings et al. 2014). Sometimes the humans take the lead, sometimes the computer does and this relationship can vary dynamically.

AI and computational thinking can actually be seen as two sides of the same coin. AI is about enabling the computer to solve problems we consider requires intelligence, or casually speaking, enable computers to “think.” Computational thinking turns this around and asks the question: how can people become better at solving problems by learning from how computers do?

Conclusion

Artificial intelligence is a fascinating topic, which has both a technical side rooted in computer science and mathematics and a humanistic side rooted in cognitive science and sociology. This chapter tries to provide an overview of the subject, mostly from a computer science perspective, but with several connections to social sciences. The dynamic and productive research field will most likely continue to evolve through the stimulation and interaction with many different fields of research. Even though a lot has been accomplished, the most important and fascinating breakthroughs probably lie ahead of us. AI is still a young research discipline that has much to offer to every researcher that wants to explore what it means to be human and what the enigma of intelligence really is.

References

Artificial intelligence


Wagner, D. N. (2020c): The Nature of the Artificially Intelligent Firm – An Economic Investigation into Changes that AI Brings to the Firm, *Telecommunications Policy* 44(6), 101954.
