Chapter 12
Artificial Intelligence and Algorithms: True Progress or Just Digital Alchemy?

Vincent Heuveline and Viola Stiefel

Abstract From chess computers to self-driving cars to the great science fiction successes in the media field—AI is omnipresent today. Starting from the question of the interaction between human and machine, this article first deals with the distinction between strong AI, which is primarily at home in the cinematic world, and weak AI, under which all actual AI systems fall today. The aspect of learning and the role of algorithms here are of eminent importance for the research and further development of the AI systems that exist to date. On the basis of artificial neural networks, computers learn, for example, to distinguish between images of cats and dogs. But can the AI also be given more weighty decisions? And how does the algorithm make a decision, mathematically speaking? What happens if the data the computer learns with is error-prone? The consequences of these considerations undoubtedly open up a range of new issues that are not exclusively relevant for research, but for society as a whole, and which will become increasingly central with the growing use of AI.

1 Introduction

Not infrequently, the question is raised how it can be that a computer generates new knowledge that the programmer did not intend at all and which is completely unknown to him. One should actually be able to assume that the computer slavishly executes the sequence of binary commands and instructions that the programmer defined by means of his software development. The computer does not make mistakes. At most, they occur when the software developer has overlooked a few of the famous bugs in his software, causing the machine to behave erroneously or even erratically. In this style, the relationship between human and machine is clearly

V. Heuveline (✉)
IWR - Engineering Mathematics and Computing Lab (EMCL), University of Heidelberg, Heidelberg, Germany
e-mail: vincent.heuveline@uni-heidelberg.de

V. Stiefel
University Computing Center, University of Heidelberg, Heidelberg, Germany
e-mail: viola.stiefel@urz.uni-heidelberg.de

R. M. Holm-Hadulla et al. (eds.), *Intelligence - Theories and Applications*,
https://doi.org/10.1007/978-3-031-04198-3_12

defined: the human dictates the tasks that the machine has to implement. The computer, as the executing instrument, does not critically weigh its reaction and does not show any feelings such as boredom, even if the tasks to be performed consist of repetitive, tedious steps. The tasks are expected to be performed deterministically and reproducibly. The power supply ensures that the bytes and bits always flow in an orderly fashion and according to plan within the electronic circuits. In this context, there is little room for the machine to solve unexpected or even creative tasks. Humans thus think they are always in complete control of the machine. However, this widely held view is very deceptive. For example, a navigation system can calculate the shortest route between the city of Heidelberg and the wonderful medieval town of Bad Wimpfen. However, we cannot assume that the developer of the navigation system knows all the insider tips of the explored area—in our case Heidelberg—and decisively designs routes and plans for it. Rather, the programmer will implement a procedure—generally in the sense of a mathematical algorithm—that is capable of calculating the shortest path between two points on a map. The question that arises here is whether the programmer still has control over their software. What does it mean to have control over an algorithm? Can unexpected results arise that were not initially intended? Can an algorithm, when combined with data, generate knowledge—in this case, the shortest or fastest path—that the programmer was not aware of in the first place? In this chapter, we will address, step by step, these fundamental questions, which are of an essential nature for understanding and evaluating AI. In the process, we will discover that the relationship between human and machine in this context is more subtle than initially assumed.

2 Strong Versus Weak AI

The list of science fiction authors and filmmakers who have devoted their works to the subject of AI is extremely long and varied. One almost constant in these books and films is that the AI portrayed is at least equal to human abilities in virtually all areas, if not surpassing them. The above-average reasoning ability of Commander Data from the science fiction series Star Trek, combined with his seamless encyclopedic knowledge in the sense of Big Data, makes him a fascinating character who surpasses human intelligence in almost every area. The ability to communicate in all natural known languages, inherent in the humanoid character C3PO from the movie Star Wars, is no less impressive. To humans, this cognitive superiority may seem at least respect-inspiring, at times even frightening. Stephen Hawking, for example, has always warned of the dangers posed to humanity by artificial intelligence. The ubiquitous media portrayal of AI only contributes to a limited extent to calming and defusing the situation. James Cameron, for example, sets further accents in connection with artificial intelligence in the second film of his well-known film series Terminator with the cyborg of the same name: Terminator combines the superlatives of all human abilities to achieve an overriding goal: to save mankind. In the process, Terminator even understands and masters a skill that is considered an exclusively human attribute—humor. All these characters have in common that they represent the expression of a so-called strong artificial intelligence, which encompasses all sides of human intelligence—also and especially in the combination of the different abilities (Flowers, 2019; Liu, 2021).

The quest for superhuman abilities, superpowers, and hyperintelligence has fascinated mankind since time immemorial, and for this reason is reflected not least in all forms of media. An example from Greek mythology would be Icarus, who ultimately failed because he wanted to reach too high. It remains to be seen to what extent the same fate threatens today's efforts of mankind to make itself godlike through a strong AI (think of Harari's Homo Deus) (Fjelland, 2020). At present, however, it can be stated that existing technology is far from enabling strong artificial intelligence as a reality. Today's AI systems fall under the category of weak AI ("weak artifical intelligence") (Walch, 2019): human intelligence or human cognitive abilities are only matched and possibly surpassed in delimited sub-areas. Image and speech recognition, automated translation, and self-driving cars are just a few examples of where (weak) artificial intelligence is used productively today.

3 Weak AI Is Mathematics

The Dartmouth Conference ("Dartmouth Summer Research Project on Artifical Intelligence"), which took place in 1956 at Dartmouth College in Hanover, New Hampshire (USA), is considered the beginning of the study of artificial intelligence in the sense of the concepts and approaches we use today.

The name "Artifical Intelligence" comes from the initiator of the conference, John McCarthy. In the context of this conference, Marvin Minsky, Claude Shannon, John von Neumann, and Ray Solomonoff should also be mentioned, who have had a very strong influence on the further developments of AI. A close examination of both the topics covered and the expertise represented makes it very clear that the underlying AI concepts are grounded in mathematical abstractions (Shaffi, 2020; Garrido, 2010). In the context of this spirit of optimism, the closing words of the conference seem on the one hand promising, but on the other hand still cautiously non-committal: "[…] 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" (https://250.dartmouth.edu/highlights/artificial-intelligence-ai-coined-dartmouth). An important insight for the past decades from this conclusion and from the whole conference is that a computer can perform more than just the core task of scientific computing—classical numerical simulation. A computer is indeed capable of learning using appropriate algorithms. However, experience shows that this can generally only be accomplished if one has significant computing power and training data (Deisenroth et al., 2020). Moreover, the past decades have shown that developments in computer technology have acted as a catalyst for AI development.

Moore's Law, which is actually more of an observation, states that every 18 months there is a doubling of the available computing power. This rule of thumb, which has not been disproved to date, corresponds to exponential growth. Such huge technological developments are undoubtedly crucial to the long-celebrated successes of AI in the game of chess and Go. Supercomputers and high-performance computers, which are heavily used for AI, have now become symbols of the greatest possible computing power available. The number of floating point operations per second, or FLOPS, is a measure of a computer's performance. Today, the fastest supercomputers are capable of computing around 500 PetaFLOPS (Peta = 10^{15}). This corresponds approximately to 500,000,000,000 operations per second.

Even for experts, such orders of magnitude are challenging and difficult to grasp. This concentrated computing power is a necessary but not sufficient condition for the successful use of AI. The learning procedures required in AI are based on training data, on the basis of which the computer is able to learn facts. In general, the more precise and comprehensive this data is, the better the computer can learn. A classic example of this in textbooks is distinguishing between cats and dogs on the basis of images. The computer is trained to "learn" with the help of a large number of pictures of cats and dogs. When a completely new image is presented to the computer, it is then able to distinguish whether it is a cat or a dog. The trick is that it is irrelevant whether the computer knows the exact breed of dog or cat presented. Like a human being, the computer has learned characteristics with the help of the training data, on the basis of which it can distinguish the animals—almost always correctly. It is essential that the distinguishing features have not been explicitly programmed in advance, but are automatically defined by the computer via the corresponding AI algorithms.

Experience shows that the quality and quantity of the training data are crucial in this respect. This is where technological IT developments play a crucial role. In an era of ubiquitous digital communication, but also of networking, e.g., via the Internet of Things (IoT), digital data is produced in a number of constellations. The term "Big Data" is self-evident in this respect: in several application areas, considerable amounts of structured and unstructured data are produced. It is not the data as such that is important, but the possibility of generating knowledge and insights from this data. Many AI systems thrive on the presence of such data and are thus able to extract important and new insights from it (Sun & Wang, 2017). In this context, it is significant that the amount of data has become so extensive in some cases that a single human would not be able to analyze this data without computer support, sometimes using AI-based approaches. Are humans thus losing control over their decision-making? How reliable are AI systems really? An answer to these questions requires a closer look into how AI algorithms work.

4 Algorithms for AI

The use of AI presupposes that the computer used has the ability to learn. A first—but erroneous—thought would be to transfer the responsibility of explicit learning to the programmer of the computer. The programmer, as the "chief pedagogue" of the machine, can prescribe rules and instructions for action via the developed software, according to which pattern the machine has to react in given situations. A major disadvantage of such an approach would be that the programmer himself would have to have the required knowledge in order to be able to transfer it to the appropriate software. Consequently, the programmer of the Go game would have to know all the tips and tricks of the Go game and implement them explicitly. Conceptually, this would mean that the existing knowledge of the machine would always be bound to the skills of the programmer team and thus de facto be considerably limited. Such a machine would hardly be able to defeat the world's best Go players. However, the successes of the Alpha-Go software against the Go champions of this world have provided proof to the contrary. The machine has always been able to win the ancient Chinese game of Go with clear superiority. How is that even possible?

The real trick is to teach the machine how to learn for itself, more or less according to the principle of help for self-help. The programmer remains the "chief educator" of the machine. However, they do not try to teach the machine the content that they generally neither know nor have mastered. Rather, they convey a methodology for how the machine can learn for itself. Such approaches fall under the term "machine learning": The idea is to artificially generate knowledge from digital data that enables the machine to make decisions on its own (Thesing et al., 2019). The practical implementation of such approaches is carried out with the help of algorithms developed specifically for this purpose. These algorithms can be roughly divided into two categories: supervised learning and unsupervised learning (Brownlee, 2019; Radford et al., 2015). The exact description of these different groups is beyond the scope of this paper. Accordingly, we will focus exclusively on the algorithms of the supervised learning category in the following.

In supervised learning, the computer learns from given pairs of inputs and outputs. For example, an image with a cat (input) is associated with a value of 0 (output) and an image with a dog (input) is associated with a value of 1 (output). The trick is to define a mapping or function between the inputs and outputs in such a way that even unknown images are correctly classified according to the above principle. For the definition of such mappings, the use of so-called neural networks has proven very useful (Saxton et al., 2019; Yosinki et al., 2014).

In biology, neural networks refer to structures of the brain of animals and humans. Neurons form an extremely complex network—in the human cerebral cortex 10 billion neurons work in fine coordination. Each neuron is connected with about 2000 other neurons. The ability to learn is achieved by changing the connection strengths of the existing neurons. Thus, information is not stored in individual neurons, but is represented by the entire state of the neural network with all connection strengths. In the field of AI, people use artificial neural networks which are strongly inspired by their biological counterpart (Nikolic, 2017). However, artificial neural networks are concerned with obtaining an abstraction in terms of model building, which can be used to define the mapping between inputs and outputs in the best possible way. The learning process based on such artificial neural networks consists in determining weights along the connections (edges) of a graph for which the neurons act as nodes. The pairs of inputs/outputs as training data are used to determine these weights. Mathematically, this is a model calibration in the sense of parameter identification. Here, the weights along the edges of the neural network are the parameters to be identified: the learning process as a parameter identification problem.

It should not go unmentioned that, from a mathematical point of view, there are still a number of open questions concerning the properties of such neural networks. For example, the determination of the dimensioning of such a neural network for a given application is still a challenge, which generally has to be defined empirically via numerous tests. Some digital alchemy is always necessary here. For simple neural networks, one can prove that the underlying methodology corresponds to known procedures from the field of numerical optimization. Thus, for such methods, one has the decidedly important support of mathematics, which provides a foundation for both the understanding and convergence statements of the methods. Unfortunately, for many procedures that have proven themselves in practice, there is little mathematical insight into why these procedures work and whether this is indeed always the case. This explains why this technology is repeatedly referred to as a black box model. For the use of AI in critical areas, such a situation may hold some dangers. For example, how sensitive is the neural network to erroneous data? There is still a very great need for research here so that such approaches are not confirmed by empiricism alone.

5 AI as a Black Box

In specific sub-areas, AI already surpasses human cognitive abilities. Quantities of training data are processed that a human brain could neither store nor process in an entire lifetime. As impressive as such results are, the question arises whether important or even critical decisions can be made at all on the basis of such results.

5.1 Can AI Technology Be Fundamentally Trusted to Make Important Decisions?

This question, which is often avoided in such clarity due to the celebrated successes of AI, is actually of significant importance. The terse statement that computers don't make mistakes no longer applies in these areas. For many applications, no one—not even the programmer who wrote the AI software—knows how the algorithm made its decision in the first place. This phenomenon is called the black box problem (Bleicher, 2017). In practice, modern learning algorithms seem to work for the most part. However, the fact is that these mechanisms that lead to a decision on the part of the AI are often simply not understood. In human/machine interaction, this is certainly an unprecedented paradigm shift. From a social perspective, this challenge also raises further questions:

- Who owns and understands the training algorithms or software?
- Who owns and understands the trained neural networks or AI models?

Knowing well that AI technology has reached quite a few areas of daily life, such foci are not only socially relevant but also of political importance (Ntoutsi et al., 2020; Mehrabi et al., 2019).

Assuming that AI algorithms learn optimally in a given metric, which—as of today—we cannot prove mathematically, the question remains whether the training data is at all suitable for the targeted decisions. In this context, several aspects have to be considered. In many application areas, the data originate from measurements that cannot be determined exactly in general. Measurement errors are commonplace for sensors. This now raises the question of how an AI system responds to both training and input data that may not be entirely error-free. The issue of the sensitivity of such systems with respect to fuzziness in the data is still a subject of research and has generally not been properly penetrated to date (Lim, 2020; Angwin & Larson, 2016).

Another aspect is possibly even more serious: What happens if the training data is incomplete and the AI system can only partially learn the data space? The danger of pre-programmed discrimination lurks precisely at this point. The magazine Focus of 12.10.2018 (Amazon, 2018) brought this issue to the point using the example of AI-based job application evaluations: "Artificial intelligence considers applications from women to be inferior." After (human) analysis of the entire process, it was found that the training data was predominantly from males. Thus, the AI system made an assessment based on ignorance rather than objective consideration. Unfortunately, this is not a marginal issue, but a challenge that must always be illuminated: Constant and transparent scrutiny of AI systems with regard to possible discrimination/bias is certainly a key task not only for academia but also for all core stakeholders in society (Yapo & Weiss, 2018; O'neil, 2016; Fu et al., 2020; Baer, 2019).

6 Interpretable AI as a Possible Solution

With the increasing use of AI, the question of the interpretability and explainability of AI decisions has become essential. In English, the term "Explainable Artificial Intelligence (XAI)" describes the field that aims to make artificial intelligence explainable (Arrieta et al., 2020; Linardatos et al., 2021). This involves understanding how and why decisions have been made by AI systems. The black box character of many AI systems should thus be broken (Molnar et al., 2020). Scientifically, such questions still pose a great challenge. In the case of multi-layer DeepLearning models, for example, these aspects cannot be answered based on current scientific knowledge. In the past decade, innovative concepts have emerged that open up new perspectives in this context. A distinction is made between ante-hoc and post-hoc approaches (Escalante et al., 2018; Samek et al., 2019; Rudin, 2019). The ante hoc methodology focuses on models that are per se and a priori—i.e., beforehand—interpretable. The post hoc approach investigates the extent to which black-box models can be analyzed interpretably a posteriori. These topics are still the subject of research in many application areas.

In interpersonal interaction, we cannot always explain why fellow human beings make one decision or another. Our trust that a decision is correct or not is based on a variety of factors that we have already learned in childhood in our interactions with other people. In this context, trusting an algorithm that one may not understand correctly naturally protrudes very far from the usual range of human experience. Thus, the tension between true progress and digital alchemy to which AI is subject can only be resolved if courageous, transparent, and innovative paths continue to be taken.

References

Angwin, J., & Larson, J. (2016). *Machine bias*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges towards responsible AI. *Information Fusion, 58*, 82.

Baer, T. (2019). *Understand, manage, and prevent algorithmic bias: A guide for business users and data scientists*. Apress.

Bleicher, A. (2017). *Demystifying the Black Box that is AI*. Retrieved from https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Brownlee, J. (2019). *A tour of machine learning algorithms*. Retrieved from https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Deisenroth, A., Faisal, A., & Soon, O. C. (2020). *Mathematics for machine learning*. Cambridge University Press.

Escalante, H. J., et al. (2018). *Explainable and interpretable models in computer vision and machine learning*. Springer.

Fjelland, R. (2020). Why generalized artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7*, 10.

Flowers, J. (2019). Strong and weak AI - Deweyan considerations. In *AAAI Spring Symposium*.

Fu, R., Huang, Y., & Singh, P. V. (2020). Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. INFORMS TutORials in Operations Research: 39–63.

Garrido, A. (2010). Mathematics and AI, two branches of the same tree. *Procedia - Social and Behaviorial Sciences, 2*(2), 1133.

Lim, H. (2020). *7 Types of data bias in machine learning*. Retrieved from https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18.

Liu, B. (2021). Weak AI is likely to never become strong AI, so what is its greatest value for us? *Computer Science*, arXiv:2103.15294.

Mehrabi, N., et al. (2019). A survey on bias and fairness in machine learning. *arXiv*.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning - A brief history, state-of-the-art and challenges. In *PKDD/ECML Workshops*.

Nikolic, D. (2017). Why deep neural nets cannot ever match biological intelligence and what to do about it? *International Journal of Automation and Computing, 14*, 532–541.

Ntoutsi, E., et al. (2020). *Bias in data-driven artificial intelligence systems - An introductory survey*. Wires.

O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threaten democracy*. Crown Edition.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

Samek, W., et al. (2019). *Explainable AI - Interpreting, explaining and visualizing deep learning*. Springer.

Saxton, D., Grefenstette, E., Hill, F., & Kohli, P. (2019). Analysing mathematical reasoning abilities of neural models. In *ICLR Conference*.

Shaffi. (2020). AI and mathematics. Retrieved from https://medium.com/swlh/ai-mathematics-699a9ea2a0d6

Sun, Z., & Wang, P. P. (2017). A mathematical foundation of big data. *New Mathematics and Natural Computation, 13*(2), 83.

The pitfalls of intelligence: Artificial intelligence deems applications from women inferior - Amazon must react. *Focus* 12.10.2018.

Thesing, L., Autun, V., & Hansen, A. C. (2019) What do AI algorithms actually learn - On false structures in deep learning. *arXiv*.

Walch, K. (2019). Rethinking weak vs. strong AI. Retreived from https://www.forbes.com/sites/cognitiveworld/2019/10/04/rethinking-weak-vs-strong-ai/

Yapo, A., & Weiss, J. W. (2018). *Ethical implications of bias in machine learning*. HICSS.

Yosinki, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, arXiv:1411.1792.

Vincent Heuveline (born in Paris, 1968) has been Chief Information Officer of Heidelberg University since December 2018. He is also Director of the Computing Center of Heidelberg University. As a professor, he heads the "Engineering Mathematics and Computing Lab" (EMCL) at the Interdisciplinary Center for Scientific Computing (IWR) at Heidelberg University. In addition, he leads the research group "Data Mining and Uncertainty Quantification" at the Heidelberg Institute for Theoretical Studies (HITS gGmbH). His research interests include high-performance and data-intensive computing as well as software development with a special focus on application areas in medicine. In teaching, he deals intensively with the topic of IT security and AI—in addition to scientific computing—in the context of dedicated lectures and seminars. Prof. Heuveline is a member of the program committees of numerous international conferences on high-performance computing. As an expert and contact person, he advises representatives of industry on topics of digitization, the application-oriented use of numerical simulations, Big and Smart Data, AI, and IT security in the industrial environment.

Viola Stiefel studied Romance Studies (French and Italian) and History at Heidelberg University and received her PhD in French Literature in 2018. Since March 2021, she has been working as a consultant at the Heidelberg University Computing Center, where her work includes topics at the interface between the humanities, digitization, and AI.