Philosophical Questions of the Manifestation of Natural Intelligence*

Alexandra Prisznyák

Mihály Héder:
Mesterséges intelligencia – Filozófiai kérdések, gyakorlati válaszok
(Artificial Intelligence – Philosophical Questions, Practical Answers)
Gondolat Kiadó, Budapest, 2020, p. 166
ISBN: 9789635560509

Mihály Héder is an associate professor with habilitation and technophilosopher working on the processing of natural languages, the design of semantic annotation technologies and systems, and the philosophy of artificial intelligence.

The book's train of thought is structured around philosophical issues related to artificial intelligence (AI). Similarly to the theme of modern 'bestseller' books, it discusses in detail the criticisms, philosophical and ethical issues surrounding the development of AI in different eras and the potential future of AI. The success of the book is mainly due to the topicality of the subject, namely the market implementation of AI and robots, in a philosophical approach.

The beginning of the historical development of AI to support the increasingly human work of human workers is usually traced back to the Dartmouth Artificial Intelligence conference in 1956. In a narrower circle, it is known that work on machine intelligence in the UK began nearly a decade earlier, thanks to efforts to crack the Enigma code during World War II. *Alan Turing*'s 1950 publication titled Computing Machinery and Intelligence drew attention to the potential of intelligent machines, which has been the subject of debate ever since. The imitation game known as the Turing test aims to measure the 'thinking'[1] of machines. Success in the

* The papers in this issue contain the views of the authors which are not necessarily the same as the official views of the Magyar Nemzeti Bank.

Alexandra Prisznyák: Institute for Training and Consulting in Banking Ltd., Senior Consultant and Artificial Intelligence & CBDC Program Manager; University of Pécs, PhD Candidate. Email: aprisznyak@bankarkepzo.hu

[1] Regarding the concept of thinking, Turing considers it to be ill-defined, too general and recommends the Turing test instead. Turing, however, never claims that the successful operation of machines equals thinking. The test is a three-actor imitation game involving two agents (a human and a digital computer) and an interrogator. The interrogator cannot hear or see the other two actors, can only communicate via text and attempts to determine, by asking questions of the other two participants, which is the computer. The objective of the game is to mislead the interrogator, during which both the human and machine are allowed to lie. To avoid an accidental hit in the game more interrogators and criteria are applied. Its goal is for the machine to be able to deceive more than 70 per cent of the interrogator and make them believe that it is human.

AI test proves the error of pessimists who predicted the performative failure of AI. It also raises the fundamental problem of man's inability to recognise his own species.

The first chapter analyses the work and main critical comments of a number of AI critics. In the view of *Joseph Weizenbaum*, the computer is nothing more than the embodiment of systematic mathematics, which has nothing to do with human intelligence, it merely follows algorithmic instructions written by programmers. To prove this, in 1964 he began writing a computer program to demonstrate the intellectual capacity of man and the limited capabilities of machines. With the development of the chatbot ELIZA, Weizenbaum's aim was to change the initial positive impression of the chatbot to the impression of an illusion/communication with people, as the conversation progresses (when the machine's limitations are revealed in an unexpected situation). Contrary to his intentions, Weizenbaum was shocked by the enthusiastic reception of the market. The creation of ELIZA was seen by both the professional community and the market as a key moment that could open many doors for development, including in the field of the computer processing of natural languages. For the remainder of his career, Weizenbaum sought to raise awareness of the dangers of subservience to machines and criticised AI research. Through the example of Weizenbaum, the author illustrates the belief in AI of the period, a boom era that later led to the first and second AI winters[2] and finally to the dynamically evolving digital era of today and the rise of AI.

In order to systematise the types of critical comments, the author necessarily addresses three important areas of AI critique: the performative failure, the phenomenological failure and the dystopia associated with the development of AI. The performative failure is a critique on the inability of AI solutions to work, which has occurred in many cases during the AI winters, breaking the momentum of progress. The criticisms are based on the view that machines are not capable of performing tasks that require intelligent behaviour. The basis of phenomenological failure is provided by the 'imitative nature' of AI, as mentioned above. The operation of machines differing from the mechanism (biochemical principle) of the human mind (quantum computers are a binary exception) results in a critical type according to which the experience of consciousness (such as the experience of emotions) is not available to machines (see for example Searl's Chinese Room Argument). In the context of the phenomenological failure, Turing considers the starting point of investigation based solely on the non-human nature of the operation of machines to be wrong. Accordingly, he does not claim that they are similar to human functioning and have the same experience as intelligent humans. He focuses the investigation solely on the fact that thinking is not sufficiently defined to draw such a conclusion.

[2] The AI winter marks a period when AI development results are stagnating, falling short of investor expectations, resulting in a halt in development and a public disengagement from enthusiasm for AI technology.

However, with the emergence of intelligent machines, the content of the concept necessarily requires a rethinking of the definition. Thus, the phenomenological approach should be interpreted as an arbitrary critical comment. The author draws attention to a very interesting fact in this respect: '[...] if phenomenological success is impossible, then the uploading of consciousness is the death of consciousness'. Philosopher *Hubert Dreyfus*, one of the most active representatives of AI critics, saw AI as a waste of resources. In his view, machines are only able to operate according to pre-programmed rules. Thus, a rule-based machine executing a set of instructions will never be able to demonstrate human behaviour. He backs up his standpoint by the failures of research projects carried out in the course of AI winters. He explains the persistence of the spirit of research despite this using four premises: (1) biological (the human brain processes information in discrete units, so there must be a biological element that resembles the elements of digital technology); (2) ontological (accessibility of information); (3) psychological (the information processing of the mind based on formal rules); and (4) an epistemological assumption (all knowledge can be formalised, Boolean functions). Dreyfus stuck to his opinion even after a machine (IBM Watson) beat humans for the first time in the quiz show *Jeopardy!* in 2011. Dreyfus's four premises are based on the – erroneous – assertion that both the human mind and the computer are general symbol-processing machines, and he makes a strong case for his position by questioning the formalisability of human knowledge. However, the properties of the general symbol-processing machine are typical of the Turing machine (epistemological understanding and design), and thus identifying it with that is as much a mistake as identifying a human with it. However, Turing does not identify the two in his model. Dreyfus argues that digital computers have shortcomings in terms of perception and action (which occur in the absence of a physical representation). The 'living' proof of the fallacy of the bases of his argument is the Deep Blue computer defeating world champion *Garry Kasparov*.

The problem area of what is a manifestation of natural intelligence (imitation) explores the ability of machines to be conscious. Machines, which operate without living and absorbing the internal experience (consciousness), are fundamentally different from intelligent human behaviour. *John Searle* denies the ability of machines to understand when he creates the Chinese room thought experiment. His assumption is that the ability of machines to understand is inconceivable without the simultaneous achievement of performative and phenomenological success. The counterarguments to the experience of understanding have addressed many aspects of the thought experiment, such as isolation, the limits of the agent (whether an isolated slice of the human brain is capable of understanding), extrapolation, learning, memory and others that can lead to an intellectual experience similar to understanding. As a counterargument, experts point to the fact that intelligent robots are equipped with sensors. Following on from Searle's thought experiment,

in the second chapter of the book the author details his main critique based on *Daniel C. Denett*'s 'intuition pumps' principle: the time factor.

Another dimension of the criticisms is the fear of the extent to which AI will free up manpower. Regarding the epistemic limits of AI escalation, the author stresses that AI beyond a certain point encounters a limit to the expansion of knowledge. Thus, a sudden escalation is highly unlikely. The gradual integration of AI into work is significantly determined in several ways. The friction created by optimisation, increased productivity and job creation/destruction (reallocation of tasks) can contribute to lowering the prices of goods/services and, along with this, to improving their accessibility – also for social groups previously excluded from consumption. However, the increasing use of AI is leading to labour market polarisation. Through its direct effect, well-defined tasks that can be described by rules (monotonic, routine) will be replaced, while through its indirect effect, new tasks will be created. Consequently, intensification of the human–robot interaction is inevitable, sometimes even in domains where the nature of the task would lead to automation, but the involvement of an expert and the review of the workflow require the presence of a human participant in the interaction process. In the new division of the work process, the following emerge as bottlenecks: (1) the ability to perceive and manipulate, (2) creative intelligence, and (3) social intelligence. The steady pace of AI development will largely determine how far in time we are from achieving the super-intelligence that is considered utopian in the present age. At the same time, there are also fundamental questions about the rights of artificial intelligence appearing alongside human intelligence. Thus, the mandate of intelligent AI for self-protection or protection against being shut down can now make sense.

The author illustrates the critical remarks made in the first chapter with examples of performative successes and areas of application discussed in the second chapter. It details, among other things, the successes achieved in SHRLDU (program), Shakey robot, MYCIN (program), Herbert (robot) and chess programming. The author also discusses the prominent role of neural networks in the blossoming of artificial intelligence, and then analyses the nature of computation by examining the nature of machines. Evolutionary processes are one of the most advanced types of computation today, leading to programs that can dynamically change and modify themselves. As examples, the Evolutionary Game of Life and AARON (the robotic artist) are detailed, demonstrating that if the output of the computation is accepted as an outcome product which also has non-physical nature, the examples just mentioned cannot be evaluated. In some cases (e.g. chess), the interpretation of the output may also be realised as symbolically represented data. Looking at the issue of formalisation, the author concludes that not only formalised problems can be solved by computation.

The final chapter discusses the present and future of ethics, the amoral agent and AI. Rejection of the moral capacity of artificial intelligence automatically implies rejection of its capacity to take responsibility. Consequently, companies that make, operate and own AI need to address the issues of ethical standards and the tangible (material compensation) / intangible (reputational) consequences of the redistribution of liability to be borne. In the author's view, ethical responsibility requires the joint responsibility of several market actors, including the regulator, the manufacturer, the maintainer/operator and even the consumer. Consensus on phenomenological and performative critiques could be an important milestone for further progress.

Héder's book provides a relevant starting point for approaching artificial intelligence and the related technologies from a business perspective. The book can be recommended to the community of economists and lawyers as it challenges researchers in a number of existing and emerging research areas, such as the impact of AI on productivity and the labour market, the business and ethical aspects of AI system design, the challenges of integrating AI into the organisational culture (change management), the role of visionary management, legal and technological issues of AI, the ways to measure AI-led efficiency, the impact of AI on corporate value, and other areas of interest.