IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/44AI/2024/4450124/
Vol. 2024 (22), Article ID 4450124

The AI Decision-Making Schema: Controlling the Golem*

Jacek ENGEL[1], Dariusz RAŚ[2] and Anna PRUSAK[3]

[1]Centrum Badawczo Rozwojowe EGM S.A., Poland

[2]The Pontifical University of John Paul II in Krakow, Faculty of Communication Science, Krakow, Poland

[3]Cracow University of Economics, Institute of Quality Sciences and Product Management, Krakow, Poland

Correspondence should be addressed to: Dariusz RAŚ, dariusz.ras@upjp2.edu.pl

* Presented at the 44th IBIMA International Conference, 27-28 November 2024 Granada, Spain

Abstract

The present paper has a review and conceptual nature. The initial assumption situates the controversy around AI within decision-making processes. According to the authors, decision-making power (or decisiveness) is the factor that should determine the use of this technology. Therefore, the considerations presented here are not only about the essence of AI in the context of sophisticated algorithms, but also about decision-making as an inherent attribute of humans, entire civilizations and economies. Historical and contemporary methods of decision support systems are reviewed, including advanced technologies such as supercomputers. Finally, the authors' concept called the "Golem Button" is presented, which aims to provide managers and legislators with a tool to structure standards and guidelines on how artificial intelligence systems (contemporary and future) can be used safely in different areas of the economy.

Keywords: AI, decision-making, autonomy, algorithms, golem

Introduction: around the definition of AI

Artificial intelligence (AI) has multiple meanings; however, there is no generally accepted definition of this term (Schuett 2019). On the other hand, when considering the role of AI in decision-making processes, it would be appropriate to start with an overview of knowledge concerning both AI and decision-making processes. The review of literature shows that these terms are closely intertwined, and that any discussion on the role of AI in various areas of human life boils down to simple questions: To what extent, and where, can humans transfer their decision-making to AI systems? As regards the already transferred decision-making to an AI system (machine), are humans able to responsibly control it?

Cite this Article as: Jacek ENGEL, Dariusz RAŚ and Anna PRUSAK, Vol. 2024 (22) "The AI Decision-Making Schema: Controlling the Golem " Communications of International Proceedings, Vol. 2024 (22), Article ID 4450124, https://doi.org/10.5171/2024.4450124

It is impossible to quote all definitions of AI, so only the most characteristic and most frequently quoted ones will be provided here. The term "artificial intelligence" was first used in 1956 by the American computer scientist John McCarthy. He defined AI as , "the science and engineering of making intelligent machines", with "intelligence" meaning "the computational part of the ability to achieve goals in the world" (McCarthy 2007). One aspect of this definition is worth noting: the reference to the source and genesis of AI and its creators, namely computer scientists (programmers). This is because AI from the technical side is a product of humans, a system of algorithms and codes.

Many research studies on AI point to the achievements of the British mathematician Alan Turing, who worked during the Second World War, among other things, to break the ciphers of the Third Reich (i.e. the Enigma machine). In 1950, he proposed the so-called Turing test to assess whether a robot can imitate human logic and behavior to such an extent that it is not possible to distinguish between the two responses. For this reason, he is considered the forerunner ("father") of artificial intelligence. He defined AI simply as "a computer that passes the Turing test" (Turing 1950).

Contemporary definitions refer to a so-called "intelligent agent", which is described as a software system that perceives its environment through sensors and interacts with that environment through actuators; while the word "intelligence" denotes the ability to choose an action to maximise a measure of performance (Russell and Norvig 2009). The latter part clearly refers to decision support; the simplest definition of a "decision" defines it as one of the possible options in a given decision problem, while decision-making is the process of selecting that option (Prusak and Stefanów 2014).

On the technical side, decision-making can take place on both a macro and micro scale; thus, it ranges from choosing a marketing strategy or selecting a medical treatment to determining the direction of movement of a robotic arm. As far as decision-support processes are concerned, they can be described in one general phrase: identifying the course of action that most closely meets the criteria most important to the decision-maker. In the context of the involvement of any form of AI in any decision-making process, it is therefore necessary to ask oneself who is the decision-maker and what selection criteria guide him or her. The answer to this question is particularly relevant in the light of those definitions of AI that describe it as "a system that can do no worse than a human being in any world" (Dobrev 2005).

Ethicality should be the primary measure of the criteria in decision-making. Although AI is a specialisation of computer science, robotics, technology, discussions and research in this area are shifting to the social sciences and humanities. For example, the recent European "White Paper on Artificial Intelligence", on the one hand uses an "engineering" definition of AI, which according is "a collection of technologies combining data, algorithms and computing power", and on the other hand indicates its potential to impact on social life (European Commission 2020).

In addition to its benefits, the technology generates a number of key challenges for the well-being of humanity, related to privacy, algorithm transparency, bias, accountability, labour market impact, military applications and much more (Skrzek 2024). For example, advanced claims processing algorithms are used in the insurance industry. One of the main challenges is to ensure that learning systems do not prioritise decision criteria leading to the perpetuation of social inequalities (msm.co.uk 2024).

For the aforementioned reasons, AI has become the subject of legal deliberations, with numerous publications also addressing its legal definitions and the regulation of the use of AI in different countries (Schuett 2019). The aforementioned White Paper calls for regulation in order to "develop trustworthy artificial intelligence to ensure a coherent legal framework, the trust of citizens and to use the potential of this technology in the transformation of the EU digital economy based on data and its applications" (European Commission 2020).

The aim of this paper is to show that the main challenge associated with the "symbiosis" of humans and AI is to maintain the "decision-making balance", i.e. the balance between assisting humans in making decisions and making decisions instead of them, i.e. the full autonomy of AI. The latter - in conjunction with the "humanisation" of AI - would mean a loss of human control over key decision-making processes, which in practice could lead to consequences so far beyond imagination that their nature and scope currently remain the domain of *science fiction*. The concept of "Golem button" was developed to help identify the areas in which humans should maintain full control over decision processes.

The research method used in this article is a review of the literature related to the application of AI in decision-making processes. Due to the specificity of the subject matter, this review will be supplemented with selected information extracted from various press reports, as this is a socially relevant topic that, in addition, has been causing an unrelenting media buzz for years.

Decision Support as An Element of Civilisation

Decision-making is an intrinsic attribute of every human being, both in private and professional life. Human beings have been dealing with decision-making since the beginning of the world, which is constituted by "free will", which is decisiveness (the right or possibility to decide on something). But with decisiveness also comes responsibility - not only for oneself, but also for others. A widely recognised moral code and a primer to support human decision-making is the Bible (Judeo-Christianity, Western civilisation). According to the Bible it is *man* - and not anyone else on earth - who is supposed to be the primary decision-maker, i.e. the subject who makes the final decision and takes responsibility for its consequences.

As a result of his own decisions, man - unlike a machine - can achieve a blessing or a curse. The measure and evaluation of human action based on these values is not only on the positive side (desirable and happy), but also on the negative side (meaningless and leading to the destruction of man). Machines, using bits in communication and programme execution, do not have this possibility. Their choices are programmed on the model of man (not God), and therefore they are "morally flattened" due to a different management system than that of humans.

Machines operate on the basis of a preset algorithm. They do not improvise in a human way, they are not "faithful" in a human way, because they do not take responsibility either for their thoughts, words or actions. Their domain is the fulfilment of commands, efficiency in the execution of tasks, and possibly the creation of new types of solutions whose fulfilment is derived from the old ones. Once a fully autonomous and self-aware AI is in place, one will not - presumably - be able to count on a decision-making process based on moral decisions, but rather on targeting actions to achieve set goals and activities within the physical world. The binary system of computing does not mimic decision-making processes based on the rules of God the Creator but is based on the acceptance - or lack thereof - in its human creator.

The act of managing machines and algorithms that will never be "free" in the human sense, will not experience sin and death, will not experience grace or life in the human sense, but will always remain soulless in this respect. Hence, there is a need to realise the limitation of AI mechanisms and to reflect on the search for a control system. The main problem with the further development of AI can therefore be described as follows: *not whether to develop AI, but how to control it*.

However, the history of civilization shows that in the decision-making processes, man often reaches for non-biblical instruments to support him in discerning which option is the "right" one for the task at hand. Since a characteristic feature of decisions is that, although they are made in the present, their effects will only be known in the future, people have been trying to know the future for centuries. Going back in time, the famous Delphic oracle - the temple of Apollo at Delphi, located on the slope of Mount Parnassus - is an example of an ancient system of decision support. The priestesses of Apollo - called Pythias - were women intoxicated by chemical compounds, most likely derived from bay leaves. According to historical accounts, in a prophetic trance they foretold the future (although usually in an ambiguous manner) and thus indicated what decisions the questioner should make. Several hundred alleged prophecies given at the Delphic oracle have survived to this day (Franus 2023).

The example of the Delphic oracle's involvement in resolving various dilemmas shows that people have always needed an instrument that would, as it were, "relieve" them of at least some of the responsibility for the consequences of a particular choice. In the same way, they have always looked for a way for someone or something to do at least part of the work for them. In Judaic literature, there is the Golem, a large, man-like figure fashioned from clay by Rabbi Yehuda Low ben Bezalel of Prague (also known as the Maharal) in order to defend Jews from attacks by the populace in the second half of the 16th century. This rabbi brought the clay figure to life by means of secret rituals, and finally placed a parchment with the inscription *Emet* (Hebrew for *truth*) in the Golem's mouth. The creature was mute and mindless, had no will of its own and was therefore merely an executor of decisions made by man. This changed,

however, after defending the Jews from attacks: The Golem became enraged and began to murder those it was supposed to serve. According to one version of the legend, the Maharal took a parchment from the clay creature's mouth and crossed out the letter “E” from the word *Emet*, leaving only Met (Hebrew for death). This resulted in the death of the Golem (Waincetel 2020).

It is worth noting that the Golem figure from the above legend is often used as an allegory for AI, or rather the loss of human control over it. This is tantamount to a loss of decision-making power; whereas the Pythias merely advised decision-makers without making decisions themselves, the Golem has transformed into a decision-maker whose priority decision-making criteria contradicted the criteria and rules of life of the community for the benefit of which it was created, after all.

Modern science and governments from around the world try to establish a legal order to prevent the uncontrolled development of AI. Just as the legendary Maharal replaced the piece of paper in the Golem's mouth in time, man is trying to prevent a situation where he would lose control over something that he himself, after all, created for his own good and benefit. Judeo-Christian, ancient (Greek) and Kabbalistic (Jewish) narratives all point to a certain convergence: the key issue in the process of settling the records of the law is the preservation of decisiveness, i.e. the ability to decide for oneself and one's life, without condemning to dependence on choices made by machines. The limit of the development of AI is therefore human decision-making power - and this applies to both big issues like human life as well as to minor ones.

Rationality and Multi-Criteria Decision Support Methods

The aim of this part of the article is to present algorithms for selecting the optimal decision option, i.e. the decision support process, in a synthetic way. Before doing so, however, it is necessary to clarify some additional concepts and make some important remarks. The methods and algorithms referred to in this section do not belong to artificial intelligence (although their users sometimes use sophisticated software), however, they allow us to understand the essence of procedures supporting more or less complex decisions. They can be said to be an intermediate link between ancient prophecy and advanced AI algorithms. However, it is impossible to ignore their role in considering the history of decision support, especially as they are still widely used in various areas of socio-economic life.

The act of decision-making can be briefly defined as the process of consciously and non-randomly choosing one of at least two possible decision options (Szarfenberg 2002). Even very simple decisions require at least two options (even if it is a “yes-no” choice), and each of them is always characterised by different features. However, when these possibilities (options) are numerous and their selection depends on a large number of factors (criteria), we are talking about complex and multi-criteria decisions, which require appropriate support. The support mechanism consists of recommending or singling out the option that is optimal in terms of the decision-maker's objectives and expectations (Roy 2005). This is because every decision has certain consequences; some are positive (benefits, opportunities), others negative (costs, risks) and still others neutral or uncertain about their impact. Thus, the essence of choice is to determine those options that have an appropriate (acceptable) relationship between positive and negative consequences. This type of decision is referred to as rational (Prusak and Stefanów 2014). This consideration allows us to address the role of artificial intelligence: if AI is to have decision-making at all, it must be rational, in the specific sense of the word.

For centuries, man has sought to rationalise decisions. The beginnings of a systematic approach to decision-making can be seen as far back as the 13th century, exemplified by the works of Ramon Llul (1232-1315), a Catalan Franciscan Tertiary, missionary, philosopher and theologian, and a blessed of the Roman Catholic Church. It is Llul, who lived 700 years before A. Turing, and who is considered to be the original inventor of combinatorics, used in the construction of the logic machine. He is therefore often cited as the “forefather” of AI (along with Turing). In the context of decision support, he developed, among other things, the foundations for the theory of social choice and group decisions, and was also the author of the pairwise comparison method. His work was later continued by, among others, the Marquis de Condorcet (Colomer 2013). At this point, it should be emphasised that reflections on the problems of social choice theory can be found as far back as ancient times. Indeed, ancient reflections on decision support are not only the aforementioned oracle. Group decision-making procedures and the possibility of manipulating them were already discussed by Pliny the Younger (Lisowski 2010).

A more structured approach to decision-making was observed much later, in the form of the probability calculus of winning at dice (e.g. Galileo: Reflections on the Game of Dice, published in 1718). In contrast, Benjamin Franklin's 1772 letter to the English philosopher Joseph Priestley is considered to be the basis of modern multi-criteria decision support methods. He had previously asked for advice on a certain decision, in response to which Franklin wrote back that one should first of all consider all aspects of a given decision problem, determining their relative importance. Franklin used terms such as “decision objectives”, “criteria” and “estimation of weights” (Labaree 1956).

Modern decision support methods were proposed relatively late in the second half of the 20th century. They were a response to the challenges of the modern economy and the increasing demands of the political and business world. These methods are collectively known as Multiple-Criteria Decision Analysis (MCDA) and/or Multiple-Criteria Decision Making (MCDM), and formally belong to operations research, supporting the subjective evaluation of criteria by decision-makers (Mardani et al. 2015). Most of them were proposed in the 1970s and 1980s. The most commonly mentioned in the literature are, for example, SAW (Simple Additive Weighting Method), SMART (Simple Multi-Attribute Ranking Technique), TOPSIS ((Technique for Order of Preference by Similarity to Ideal Solution) and AHP/ANP (Analytic Hierarchy Process/Analytic Network Process) (Trzaskalik 2014). These methods are not part of AI, however, they help to understand the essence of decision-making processes and represent an important link in the evolution of advanced AI systems adapted to analyse huge amounts of data.

AI as the Advanced Decision Support Systems

In order to discuss contemporary AI systems from a decision support perspective, it is important to remember that AI is a very capacious term, under which a number of advanced technologies are covered. These include machine learning (ML), deep learning (DL), neural networks (NN), robotics and many others. Artificial intelligence (AI) “uses” these technologies to mimic human behaviour, solve problems and support complex decision-making processes - which is essentially its primary goal (mindbox 2023).

Machine learning (ML) is a “subset” of AI that uses data and algorithms to mimic the human learning process. Everyday examples of the use of this form of AI are recommendation functions on the internet (e.g. shopping), while in industry and business it is the so-called Cognitive Robotic Process Automation (cRPA). It consists of improving the quality of “human” decisions in robotic processes through the use of artificial intelligence algorithms. Unlike mechanisation or automation, which aims to relieve humans of the burden of heavy, routine and relatively simple tasks, cRPA combines a tool with the ability to learn on its own, thus enabling even highly complex tasks to be performed with precision. Such a process is only possible thanks to the rapid processing of huge amounts of data (mindbox 2021).

The cRPA processes are not only used in production processes, but also in financial markets. Daniel Kahneman, Nobel Prize winner in economics, shows an example of the low efficiency of stockbrokers. No technical analysis or data visualisation tools can guarantee the kind of decision-making results that ML provides by rapidly processing huge amounts of data and taking into account correlations imperceptible to the human mind (Kahneman 2011).

As far as machine learning is concerned, three main learning methods can be found in the industry literature: supervised (algorithms learn on pre-labelled sets), unsupervised (human presence minimised, algorithms detect patterns by themselves), semi-supervised (algorithms learn on labelled data and then start to operate unsupervised), and reinforcement learning (similar to unsupervised learning, on a trial-and-error basis). As can be seen, ML uses a typical decision-making mechanism consisting of three groups of processes: 1) the decision-making process, in which the algorithm , ‘guesses’ what the pattern of action is; 2) error estimation, i.e. the extent to which the pattern matches other already known examples (if available); the key question is whether the decision pattern is correct, and if not, how far short of the ideal is it?; 3) optimization, i.e. matching the pattern to the “ideal” (Berkeley 2020).

A component of ML algorithms are neural networks (NNs), which mimic the workings of the human mind with advanced algorithms. They are used to work with unstructured data and are the basis for deep learning (DL) algorithms (“a subset” of ML). Examples of such algorithms are all kinds of web bots and digital assistants like Siri, which also use natural language processing (NLP) algorithms (Grieve 2023).

One of the most important optimisation techniques used in ML - alongside neural networks - is the so-called genetic algorithm (GE). It was developed in 1975 by John Holland and is based on Darwin's theory of natural selection, which occurs in nature. It involves searching through a space of alternatives (i.e. populations of individuals). A first population is formed at random and then the algorithm works according to the following scheme: 1) evaluation of the solutions (individuals) that make up the population; 2) selection of those solutions (chromosomes from specific individuals) that will pass on their 'genes' to the next generation; 3) the selected chromosomes undergo genetic operations and a new population, i.e. the next generation of individuals, is thus formed. However, it should be borne in mind that GEs belong to heuristic algorithms, thus providing no guarantee of finding an optimal solution (Rożek 2022).

Another ML technique is so-called transfer learning (TL), which involves transferring the knowledge gained from solving a given problem to another. These algorithms involve using pre-defined models and adding layers specific to the new task. This significantly reduces the time needed to train neural networks, which is why TL is considered a key aspect of modern AI, allowing for image recognition and processing, natural language, or autonomous driving systems (Kozon 2023).

Classification of Contemporary AI Systems

Due to the vastness of the topic, it is impossible to list all applications of AI in decision-making processes, nor to provide a "one-size-fits-all" classification. One such categorisation is the division of AI in terms of functionality, in which four categories can be distinguished, showing at the same time its gradual evolution and successive levels of development: 1) reactive AI, 2) AI with limited memory, 3) AI theory of mind, and 4) AI self-awareness (AI 2024).

Reactive AI (reactive machines) is essentially the basic and primary form of AI, characterised by the fact that it cannot use any past information in its algorithms because it has no memory. An example of reactive AI was the prototype IBM supercomputer with which Garry Kasparov lost at chess in the 1990s, or, before that, Arthur Samuel's checkers algorithm. AI with limited memory includes systems that are capable of using past experience to make decisions. Most contemporary applications of AI can fall into this category, from intelligent autonomous systems in manufacturing or vehicles, to chatbots, ChatGPT or even supercomputers. Examples of the capabilities of the latter will be discussed below. AI theory of mind is a type of AI that does not yet exist in its proper form, but which will be able to fully read and understand human emotions and expectations, as well as interact in various social interactions. Although there are currently attempts to develop robots that provide "emotionally supportive" assistance to elderly and lonely people (e.g. Kostrzewski and Patera 2024), they are not able to fully recognize their complex emotions. The fourth and highest level of AI development is self-awareness, i.e. a system with its own super-intelligent consciousness that is sentient and capable of reflection. For the time being, this kind of AI does not yet officially exist. However, there was a high-profile case of one Google employee being fired for declaring that one of its AI systems had gained consciousness (komputerswiat.pl 2023).

An interesting case in considering the role of AI in decision-making is so-called supercomputers. This is a term as broad as AI, describing devices that far exceed the capabilities of home computers, mainly in terms of computing power. Most often they are huge installations the size of football fields, consuming large amounts of power and emitting large amounts of heat. They are capable of speeding up many data processing processes, which is why these types of devices are used in the military sector or in scientific research, although their capabilities are much broader. There are currently a large number of supercomputers in the world, and the most efficient ones include Frontier, Fugaku, LUMI, Summit or Sierra (Olszewski 2022).

IBM Watson can be cited as an example of a supercomputer used in decision-making processes. Although more powerful supercomputers exist today, this one is probably one of the best known in the world. IBM began research into developing such a device as early as the 1950s, and it can be said to be one of the prototypical experiments in the area of artificial intelligence. The first supercomputer - Deep Blue - defeated chess grandmaster Garry Kasparov in the 1990s. In 2011, an improved version of Watson won the US quiz show Jeopardy! This was made possible by quickly comparing all possible alternatives in terms of probability of correctness (IBM 2024). Watson is currently available to the public and has been used in Poland since 2017 in oncology facilities (*Watson for Oncology*). This artificial intelligence has been developed specifically to assist doctors in decision-making in the area of selecting the appropriate therapy, based not only on the medical information of specific patients, but using a much broader data

resource (PAP/Health Market 2017). However, it should be emphasised that the results of data processing by supercomputers, which give even the most reliable results, are only a suggestion of a decision; it is the human doctor who is the ultimate decision-maker, taking responsibility for choosing one therapeutic path and not another. The consequences of shifting decision-making to AI are easy to imagine in this case.

The next step in supercomputer development is quantum computers. These are expected to replace today's silicon computers in the future. These computers calculate not with bits (0-1), but with qubits (quantum bits), which can take on several different values at the same time. As a result, quantum computers will be able to process huge amounts of data simultaneously and will be able to model extremely complex physical and biochemical phenomena. Although this technology is very promising, its capabilities are currently very limited. For example, qubits only work under certain conditions, such as high vacuum and very low temperatures, and are also susceptible to external disturbances. As a result, modern quantum machines take up a lot of space and look more like a chemical laboratory than a computer (Stradowski 2023).

But this could change rapidly. The almost limitless computing power of quantum supercomputers could accelerate the development of AI and lead it to rapidly reach the next stages: theory of mind and self-awareness. In any case, the postulate of maintaining control over AI decision-making should be borne in mind, regardless of the level of sophistication of the technology.

The Golem Button Concept

The simplest conclusion would be that any transfer of decisions from humans to AI is potentially risky and against the norms already presented in the biblical accounts. However, such a conclusion is not the intention of this review, and the authors are of the opinion that AI decision-making should be differentiated and graded. Hence, the authors developed their own concept called the “Golem Button” (Fig. 1), from the symbolism of the Golem presented earlier, which is an allegory of technology that is out of control.

The proposed concept describes three paths of AI-human symbiosis and the degree of human involvement in decision-making processes (“decision tunnels”) involving AI. The green path (*Approval Path*) reflects morally “neutral” decisions made by the technology that in principle do not require human approval. Examples of such decisions are autonomous systems used, for example, in vehicles, based on objective optimization algorithms. The correctness of such decisions depends on the correctness of the algorithms themselves, over which, however, humans have full control (outside the “decision tunnel”).

The yellow path (*Check Path*) shows a situation where a decision should be subject to human acceptance and ‘human scrutiny’, as it relates to a situation where a person is judged in some way, suffering moral or psychological costs. Examples of decisions in this category include learning systems for credit processing or insurance claims. The controversial case of Father Justin is also an example of moral harm to humans from AI. A Catholic platform provided a bot, posing as a Catholic priest (named Justin and dressed in clergy robes), to provide answers on the topic of faith and resolve moral dilemmas. The bot quickly got out of hand, as the answers given often contradicted the official interpretation of the Church (Chaturvedi 2024).

The red path (*Golem Button*) involves critical decisions that have a significant impact on human destiny. Defence decisions are an example. In such cases, humans must always remain absolutely decisive, in control and ready to intervene by immediately stopping any action taken by algorithmic systems that could lead to negative consequences for humans (and humanity). The concept presented here aims to provide managers and legislators with a tool (blueprint) to structure standards and guidelines on how artificial intelligence systems can be used safely in different areas of the economy.

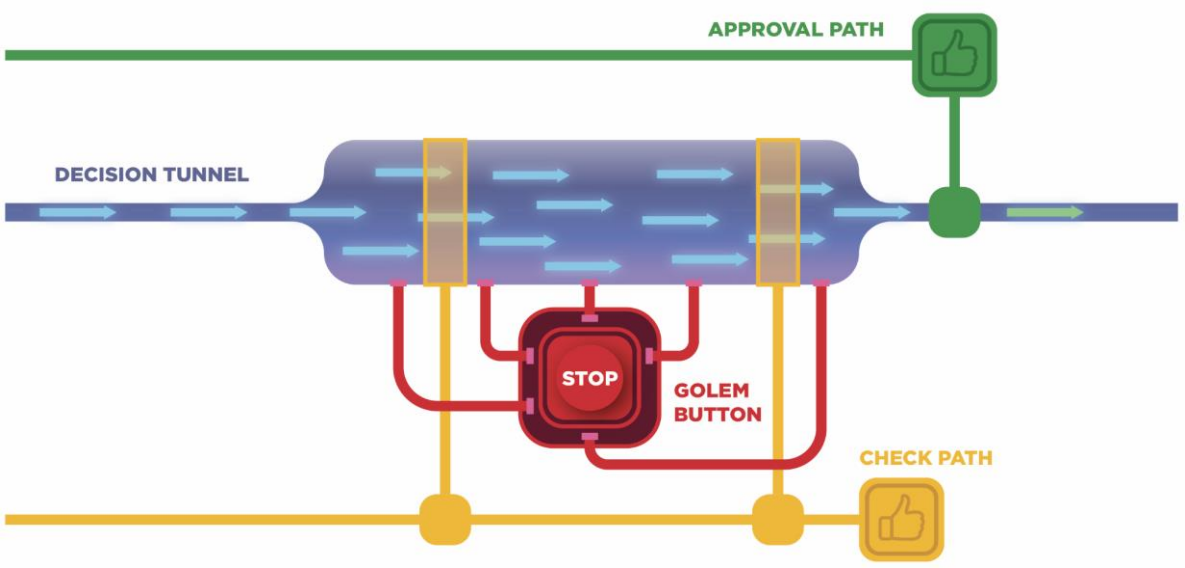

Fig. 1. The concept of Golem Button

source: own elaboration

Conclusions

Decision-making is an intrinsic attribute of every human being, whether in private, social or even professional life. However, this process is increasingly becoming a 'participation' of man-made machines. Man has been dealing with decision-making since the beginning of the world's existence, which is determined by his capacity commonly referred to as free will, which is decision-making power. Only man has the right or the possibility to "fully" decide about something and someone. But what if he gives this right partially or completely to machines built by humans? The analysis considered the practical aspects of this problem.

The intrinsic problems of human decision-making impossible to transfer to the most advanced AI technology are undoubtedly free will, conscience and responsibility. They are therefore accompanied by a broader paradigm - a set of concepts and theories that form the scientific basis in the ethical/moral, legal, educational, economic etc. fields. Each of these terms raises reflections on the facilitations that come from automation in everyday life and the consequences of their implementation in an unregulated manner. Deregulation in the case of AI autonomy fails as a matter of principle and as a matter of safe achievement of the goals of the action (the object of the act). Good intentions (the goal of the acting subject) and favourable circumstances (increasing technological capabilities) are no guarantees for the safety of humans.

The authors conclude that, in view of the socio-political tension and increasing feelings of insecurity about AI, science should accelerate interdisciplinary analytical work on the processes within the area of AI autonomy. The proposed "Golem Button" concept embeds these views and places the discourse on AI in an ethical box between limited assistance and full autonomy. With this, the authors wish to contribute to the numerous ongoing discussions in the world on legislation for the application of AI, pointing out that one of their primary determinants - regardless of the advancement of the technology - should be human-maintained decision-making.

References

- Berkeley School of Information (2020). What Is Machine Learning (ML)? Artykuł online: https://ischoolonline.berkeley.edu/blog/what-is-machine-learning/ (data dostępu: 15.07.2024 r.).
- Chaturvedi, A. (2024). AI Priest "Father Justin" Defrocked For Giving Wrong Answers. NDTV. [Online], [Retrieved November 3, 2024], https://www.ndtv.com/feature/ai-priest-father-justin-defrocked-for-giving-wrong-answers-5554764

- Colomer, Josep (2013). Ramon Llull: From 'Ars electionis' to social choice theory. Social Choice and Welfare, 40(2): 317-328. DOI: 10.1007/s00355-011-0598-2
- Dobrev, Dimiter (2005). A Definition of Artificial Intelligence. Mathematica Balkanica 19, New Series, Fasc. 1-2, 67-74.
- Franus, Agnieszka (2023). Delfy – najważniejsza wyrocznia starożytności. Artykuł online: https://www.national-geographic.pl/artykul/delfy-najwazniejsza-wyrocznia-starozytnosci#kim-byla-pytia (data dostępu: 15.07.2024 r.).
- Galileo (1718). *Reflections on the game of dice*.
- Grieve, Patrick (2023). Deep learning vs. machine learning. Artykuł online: https://www.zendesk.com/blog/machine-learning-and-deep-learning/ (data dostępu: 15.07.2024 r.).
- ibm.com (2024). IBM Watson to watsonx. Artykuł online: https://www.ibm.com/watson (data dostępu: 15.07.2024 r.).
- Kahneman, Daniel (2011). Pułapki myślenia. O myśleniu szybkim i wolnym, Wydawnictwo Media Rodzina.
- European Commission (2020). White paper on artificial intelligence. Brussels, 19.02.2020 r. https://www.gov.pl/web/ia/biala-ksiega-w-sprawie-sztucznej-inteligencji (data dostępu: 15.07.2024 r.).
- komputerswiat.pl (2023). Sztuczna inteligencja i świadomość? Czy to możliwe? Maszyna nabrała naukowca. Artykuł online: https://www.komputerswiat.pl/artykuly/redakcyjne/sztuczna-inteligencja-i-swiadomosc-czy-to-mozliwe-maszyna-nabrala-naukowca/cwmlfpn (data dostępu: 15.07.2024 r.).
- Kostrzewski, Kacper; Patera, Jakub (2024). Korea Południowa wprowadza lalki AI do opieki nad starszymi ludźmi. Artykuł online: https://www.aibuzz.pl/p/korea-poudniowa-wprowadza-lalki-ai-opieki-nad-starszymi-ludmi-ai-buzz, (data dostępu: 15.07.2024 r.).
- Kozon, Tomasz (2023). Co to jest Transfer Learning? Artykuł online: https://boringowl.io/blog/co-to-jest-transfer-learning (data dostępu: 15.07.2024 r.).
- Labaree L.W. (red.) (1956), Mr Franklin: A Selection from His Personal Letters, New Haven, CT, [za:] http://www.procon.org/view.background-resource.php?resourceID=1474 (data dostępu: 14 marca 2014).
- Lisowski, Grzegorz (2010). Uzasadnienia metod wyboru społecznego. Decyzje, 14/2010: 14-32.
- Mardani, Abbas, Jusoh, Ahmad, Zavadskas, Edmundas (2015) Fuzzy multiple criteria decision-making techniques and applications – Two decades review from 1994 to 2014. Expert Systems with Applications 42: 4126-4148
- McCarthy, John (2007). What is Artificial Intelligence? Stanford University.
- mindbox (2021). cRPA – Kognitywna Robotyzacja Procesów Biznesowych – co warto o niej wiedzieć i gdzie wykorzystać? Artykuł online: https://mindboxgroup.com/pl/crpa-kognitywna-robotyzacja-procesow-biznesowych-co-warto-o-niej-wiedziec-i-gdzie-wykorzystac/ (data dostępu: 15.07.2024 r.).
- mindbox (2023). Artificial intelligence, machine learning i neural network — co je łączy, a co dzieli? Artykuł online: https://mindboxgroup.com/pl/artificial-intelligence-machine-learning-i-neural-network-co-je-laczy-a-co-dzieli/ (data dostępu: 15.07.2024 r.).
- msm.pl (2024). Jak sztuczna inteligencja rewolucjonizuje branżę ubezpieczeń? Artykuł internetowy: https://msm.pl/blog/jak-sztuczna-inteligencja-rewolucjonizuje-branze-ubezpieczen.html (data dostępu: 15.07.2024 r.).
- Olszewski, Daniel (2022). Najwydajniejsze superkomputery na świecie -- lokalizacje i funkcje. Artykuł online: https://www.computerworld.pl/news/Najwydajniejsze-superkomputery-na-swiecie-lokalizacje-i-funkcje,441394.html, (data dostępu: 15.07.2024 r.).
- PAP/Rynek Zdrowia (2017). Komputer Watson w co drugiej diagnozie był zgodny z onkologami. Artykuł online: https://www.rynekzdrowia.pl/Serwis-Onkologia/Komputer-Watson-w-co-drugiej-diagnozie-byl-zgodny-z-onkologami,179135,1013.html, (data dostępu: 15.07.2024 r.).
- Prusak A., Stefanów P. (2014). *AHP – analityczny proces hierarchiczny. Budowa i analiza modeli decyzyjnych krok po kroku [AHP - analytic hierarchy process. A step by step construction and analysis of decision models]*, Wydawnictwo C.H. Beck, Warszawa.
- Roy Bernard (2005). Paradigms and challenges [w:] Multiple-Criteria Decision Analysis, red. J. Figueira i in., Nowy Jork.
- Rożek, Piotr (2022). Algorytm genetyczny. Artykuł online: https://fingerprints.digital/strefa-eksperta/algorytm-genetyczny/ (data dostępu: 15.07.2024 r.).
- Russell, Stuart and Norvig, Peter (2009). Artificial Intelligence: A Modern Approach. Pearson Education, 3rd edition.
- Schuett, Jonas (2019). Defining the scope of AI regulations. Law, Innovation and Technology, 15(1): 60-82. DOI: 10.1080/17579961.2023.2184135.

- Turing, Alan M. (1950). Computing Machinery and Intelligence. Mind, 59: 433-460.
- SI (2024). Sztuczna inteligencja. Słownik online: https://www.sztucznainteligencja.org.pl/definicja/sztuczna-inteligencja/ (data dostępu: 15.07.2024 r.).
- Skrzek, Kinga (2024). Etyka w sztucznej inteligencji: jak zapewnić, że AI będzie służyć dobru ludzkości? Artykuł internetowy: https://przemyslprzyszlosci.gov.pl/etyka-w-sztucznej-inteligencji-jak-zapewnic-ze-ai-bedzie-sluzyc-dobru-ludzkosci/ (data dostępu: 15.07.2024 r.).
- Stradowski, Jan (2023). Komputer kwantowy pozwoli wykonywać bardzo złożone obliczenia. Do czego nam się przyda? Artykuł online: https://www.national-geographic.pl/artykul/komputer-kwantowy, (data dostępu: 15.07.2024 r.).
- Szarfenberg, Ryszard. (2002), Racjonalność decyzji w polityce społecznej, referat wygłoszony na konferencji WDiNP w 2002 roku, http://rszarf.ips.uw.edu.pl/pdf/refwdinp.pdf, (data dostępu: 15.07.2024 r.).
- Trzaskalik, Tadeusz (2014). Wielokryterialne wspomaganie decyzji. Przegląd metod i zastosowań, Organizacja i Zarządzanie 1921(74): 239-263.
- Waincetel, Marcin (2020). Czy nieśmiertelny Golem, symbol czesko-żydowskiej kultury, narodził się w Polsce? Artykuł online: https://twojahistoria.pl/2020/07/24/czy-niesmiertelny-golem-symbol-czesko-zydowskiej-kultury-narodzil-sie-w-polsce/ (data dostępu: 15.07.2024 r.).