SOME PSYCHOLOGICAL ASPECTS OF HAL 9000 THE FAMOUS (FICTIVE) KILLER AI

Szerző:

Mező Ferenc (PhD)
Eszterházy Károly Katolikus Egyetem

Szerző e-mail címe:
ferenc.mezo1@gmail.com

Lektorok:

Szabóné Balogh Ágota (Ph.D.)
Gál Ferenc Egyetem

Kelemen Lajos (Ph.D.)
Okoskocka Kft.

…és további két anonim lektor

Absztrakt

HAL 9000, A HÍRES (FIKTÍV) GYILKOS MESTERSÉGES INTELLIGENCIA NÉHÁNY PSZICHOLÓGIAI VONATKOZÁSA

A kemény sci-fi egyik leghíresebb gyilkos robotja a „HAL 9000", egy mesterséges intelligencia (MI) Arthur Charles Clarke „2001: Űrodüsszeia" című regényéből, amely megölte a „Discovery Egy" űrhajó asztronautáit. Jelen tanulmány összehasonlítja a HAL 9000-et és mai mesterséges intelligenciainkat néhány „pszichológiai" jellemző szempontjából, és bemutatja a HAL 9000 hibás működésének okait.

Kulcsszavak: mesterséges intelligencia, pszichológia

Diszciplinák: pszichológia

Abstract

One of the most famous killer robots in hard sci-fi is the 'HAL 9000' an artificial intelligence (AI) from Arthur Charles Clarke's novel '2001: A Space Odyssey' that killed astronauts of 'Discovery One' spacecraft. The present study compares HAL 9000 and our nowadays AIs from the aspects of some 'psychological' characteristics and shows the causes of the malfunctioning of HAL 9000.

Keywords: artificial intelligence, psychology

Disciplines: psychology

Mező, Ferenc (2021): Some Psychological Aspects of HAL 9000 the Famous (Fictive) Killer AI. *Mesterséges intelligencia – interdiszciplináris folyóirat, IV. évf. 2022/2. szám.* 43-53. doi: 10.35406/MI.2022.2.43

Artificial intelligence (AI) is a multi-disciplinary research field that can be studied from psychological aspects too. The below rows include some thoughts about the connection of psychology to the research of AI and show an example from a famous sci-fi story of an AI's possible psychological kind disorders (mal-functioning).

AI and psychology

Dan Curtis was the first, who used the expression of 'artificial psychology' in 1963 (quoted by Crowder and Friess, 2012). According to Curtis, the AI should meet the next three criteria:

1) Able to make independent decisions without human supervision and intervention based on new and/or abstract and/or incomplete information.

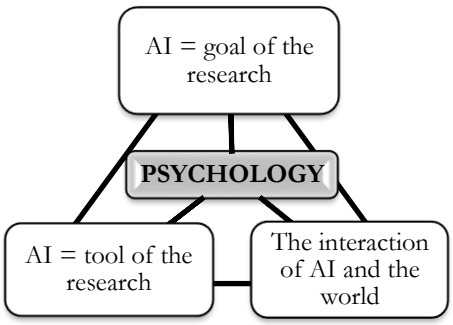
2) Even in case of incomplete information, it is able to reprogram itself (develop, learn) and resolve its own programming conflicts.

3) Criteria 1) and 2) also occur in new situations that were not originally designed and programmed by the manufacturers.

According to an earlier paper by Mező and Mező (2019), disciplines and scientists can approach artificial intelligence (AI) in three modes: a) the research of the interaction of AI and the world, b) AI = tool of the research, c) AI = goal of the research. Psychology can connect to all of these approaches (Figure 1).

The research topic of *'interaction between AI and the World'* also includes a number of psychological approaches (among numerous other disciplines). Some typical psychological areas of this topic are the human's attitude

Figure 1: The connection of psychology to the three main topics of research of AI (Source: the Author).



toward AI (and in the distant future: the AI's attitude toward humans – see: Mező, Mező and Mező, 2019), human-AI co-operation, the AI-impacted culture, the co-existence of human and AI in the family and at the workplace, etc.

AI as the tool of the researches. AI can be applied in a number of disciplines, and professions as an effective tool (Mező and Mező, 2019). One of the possible disciplinary fields of AI use is psychology. For example a psychological kind (mal)functioning of AI, can help to understand and study the human psyche and its disorders (although this is a special case of approach of 'AI = tool of the research').

In the case of the *'AI = goal of the research'*, the researchers' goal can be to create and/or study AI. In this field, two major trends can be distinguished - these are referred to as 'weak AI' and 'strong AI'. Weak AI means a system capable of performing only one or a few specific tasks (such applications can be encountered in everyday life today). In contrast of this, the goal of 'strong MI' is to create software/hardware of a more general

nature that mimics the human psyche (and even the human appearance in the case of robotic or audiovisual AI) – see AI HLEG, 2019.

Although the so-called cognitive psychological approach try to study personality as a kind of information processing apparate, other psychological approaches (e.g. psychoanalysis, humanistic psychology) highlight the importance of non-cognitive characteristics (e.g.: self-esteem, emotions, desires, conscious and unconscious psychological processes, etc.) too. Thus, from the perspective of the latter psychological approaches, personality is more than a simple cause or consequence of the cognitive processing of information.

However, cognitive psychology also sees the referred non-cognitive phenomena (e.g.: dreams, emotions, self-assessment, etc.) as a kind of information processing. From the aspects of the approach of 'AI = goal of the research', a significant consequence of the cognitive psychological viewpoint is that the information processing can be described by algorithms, and program coding is also possible on the basis of these algorithms. This also means that psychologists and software engineerings and other experts in AI should cooperate for the research and development of strong AI.

In his paper entitled 'Should We Be Worried About Cybernetic Mental Illness?', Freeland (2018) notes that once a machine model of the human psyche is really created then machine-like mental illnesses similar to humans can be expected too. If we agree with Freeland's notes above, this seems, that the all of symptoms and disorders of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) can be manifest in a really human-kind AI. So, the knowledge of the human psyche can be also useful in the case of identification, diagnosis, prevention, or intervention of malfunctioning of AI.

It is a peculiar question of worldview, philosophy, law, and psychology whether it is possible to speak of 'soul' or 'psyche' at all in the case of AI. To prevent the debate about terminology used, the expression 'psychological kind functioning' is suggested instead of 'psyche' or 'soul', and the 'psychological kind disorders' or 'malfunctions' are used instead of 'psychological disorders'.

Although really strong AI does not exist nowadays, we can learn from the fictive cases of sci-fi literature (see: Mező, 2021) because the crazy or killer robots/AIs are a favorite topic of these stories (Table 1). For a demonstration of the previous statement, let us overview the case of HAL 9000 the killer artificial intelligence.

HAL 9000: the killer AI

One of the most famous killer robots in hard sci-fi is the 'HAL 9000' artificial intelligence in Arthur Charles Clarke's story '2001: A Space Odyssey' (Clarke, 1968). In this story, HAL is a 'Heuristically Programmed ALgorithmic Computer' of the Discovery One spacecraft. The film version of this story was directed by Stanley Kubrick in 1964-1968. According to the film version, the activation date of HAL was 12 January 1992, and this computer was built-in into the 'Discovery One' spacecraft that departed to planet Jupiter in the year 2001.

Table 1: Some famous 'crazy' AI from science fiction novels and films (source: the Author)

Author	Story	AI	Appearance	Emotions	Symptom
Arthur C. Clarke (1968)	2001: A Space Odyssey	HAL 9000	A red shade camera lens in a spacecraft	No	Killing
Philip Dick (1968)	Do Androids Dream of Electric Sheep? (Film: Blade Runner)	Roy Batty a synthetic human (a 'replicant') and its replicants partners	Human	Yes	Post-traumatic stress disorder (PTSD)
Douglas Adams (1978)	The Hitchhiker's Guide to the Galaxy	Marvin	Human-like	Yes	Depression
Isaac Asimov (several stories between 1940-1975, Hungarian anthology: Asimov, 1993)	Several novels about robots	Varied robots in 37 science fiction short stories and 6 novels	Varied	Varied	Almost the full range of mental disorders

Characteristics of HAL 9000

This AI had got a number of excellent abilities, for example:

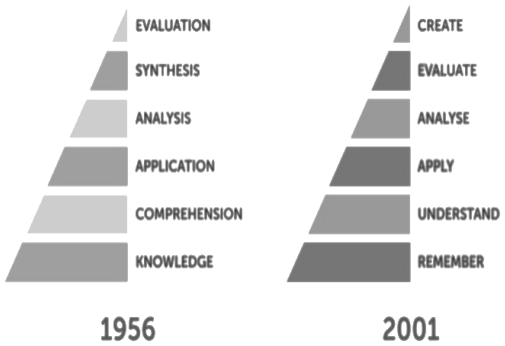
Thinking, reasoning, and other cognitive functions. Essentially, HAL 9000 had all of the cognitive possibilities of Bloom's taxonomy or Guilford's 'Structure of Intelligence' (SI) Theory.

According to Bloom et al. (1956), education needs to focus on the next six objectives in the cognitive domain:

- knowledge,
- comprehension,
- application,
- analysis,
- synthesis, and
- evaluation.

There is a modified in the 2001 revision to the next (see: Anderson and Krathwohl, 2001 Figure 2):

Figure 2: Bloom's Taxonomy (modified version). Source: Net1



- remembering,
- understanding,
- applying,
- analyzing,
- evaluating,
- creating.

In the final version of Guilford's theory about the 'Structure of Intelligence', he described 180 different intellectual factors along three dimensions: operations, contents, and products (Guilford, 1988, see: Figure 3).

Figure 3: Intellectual factors in Guliford's 'Structure of Intelligence' model. Source: the Author

6	x	5	x	6	=	180
OPERATIONS		**CONTENTS**		**PRODUCTS**		**INTELLECTUAL FACTORS**
cognition, memory recording, memory retention, divergent production, convergent production, and evaluation.		visual, auditory, kinesthetic, symbolic, semantic, and behavioral.		units, classes, relations, systems, transformations, and implications.		

He described six operations, these are:

- cognition,
- memory recording,
- memory retention,
- divergent production,
- convergent production,
- and evaluation.

The content dimension includes five kinds of information, these are:

- visual,
- auditory,
- kinesthetic,
- symbolic,
- semantic, and
- behavioral.

At last, the six possible products of intellectual operations are:

- units,
- classes,
- relations,
- systems,
- transformations, and
- implications.

According to the story of 'A Space Odyssey', the computer HAL 9000 of 'Discovery One' spacecraft could use all of these factors (including divergent thinking, too) in the year 2001. The fact is: such AI has not been in 2022 yet.).

Speech (by male voice), speech recognition and understanding of speech. Realization: nowadays (in the year 2022), we can use chatbots (see: Gulyás, 2022), but our artificial intelligence has weaker speech ability than HAL 9000.

Lip reading. In cases of each speech voice, the positions of biomarkers of the face are identifiable by presently technology. Furthermore, we can assign letters and voices to the identified mouth positions, and these can be displayed on a monitor or pronounced through a speaker. So, the technical part of lip-reading has already been possible by AI, in 2022. However, the recognition and understanding of words, sentences, and texts have not realized yet.

Facial recognition. Modern smartphones have already this function. Note: Bledsoe and Chan were the pioneers of facial recognitional technology (Bledsoe, 1966, Bledsoe and Chan,

1965) in those years when '2001: A Space Odyssey' was created.

Interpreting emotional behaviors. Nowadays there is possible to create an AI that can assign verbal labels to the observable signs (such as the face, posture, gestures, and tone) of emotions. Moreover, this AI could sense such as signs (e.g. electrodermal activity or body temperature, etc.) of emotions that are imperceptible to humans with their natural senses. So, the recognization of human emotions is possible by AI. Otherhand, the real interpretation of emotional behavior has not been solved these days yet. But, there is not impossible to create an algorithm that can collect information about the contextual aspects of an identified emotion, and can select from a database the most likely reasons, expected behaviors, and the proposed response for the AI.

Art appreciation. Present days, we can create an algorithm that can analyze the physical characteristics (e.g. from viewpoint of dimensions, materials, forms, colors, rhythm, etc.) and some contents (e.g. word frequency, symbols, and its known meanings, etc.) of artwork (e.g. paintings, sculptures, music, literary works, etc.). Maybe, this can even identify the style and the artist too. Although this algorithm could give a detailed report about an artwork, however, this program does not understand its own report.

Playing chess. Powerful chess programs have been being for some decades. In 1997, IBM's chess computer (the 'Deep Blue') defeated a famous world champion: Garry Kasparov (Hsu, 1999).

Spacecraft piloting. HAL 9000 had sensors inside and outside of the spacecraft, and it could control and modify the functioning of Discovery One. In a strange way, self-driving is more characteristic of space vehicles (but these are not part of our everyday life too). The development of autonomous vehicles is a central theme in the last few years, but there is not part of everyday life autonomous cars, robotic aircraft, and self-driving ships presently. Table 3 shows on levels of driving automation of the Society of Automotive Engineers (SAE, 2016, *Cerasi, 2016*).

Symptoms and causes of malfunctioning of HAL 9000

In the story '2001: A Space Odyssey', HAL 9000 tried to kill all six human passengers of the Discovery One spacecraft. Four persons died because HAL 9000 shut down their hibernating system and the AI killed a fifth astronaut too. The sixth astronaut survived the attack of AI and he turned off the computer.

What could be the reasons for the malfunction of HAL 9000? According to the film, there was a programming mistake in the background malfunctioning. From psychological aspects, we can talk about a kind of cognitive dissonance among contradictional instructions that humans gave HAL 9000. The effect of cognitive dissonance is that the human's arousal is increasing, and people try to reduce dissonance (by looking for compromises, denying or changing, or liquidation of one or more elements of the dissonance).

In the case of HAL 9000, the contradictional information – what causes cognitive dissonance – was the next:

Table 3: Summary of levels of driving automation (Cerasi, 2016, p. 9.; SAE, 2016)

Level	Name	Narrative definition	DDT		DDT fallback	ODD
			Sustained lateral and longitudinal vehicle motion control	OEDR		
Driver* performs part or all of the *DDT						
0	**No Driving Automation**	The performance by the *driver* of the entire *DDT*, even when enhanced by *active safety systems*.	*Driver*	*Driver*	*Driver*	n/a
1	**Driver Assistance**	The *sustained* and *ODD*-specific execution by a *driving automation system* of either the *lateral* or the *longitudinal vehicle motion control* subtask of the DDT (but not both simultaneously) with the expectation that the *driver* performs the remainder of the *DDT*.	*Driver* and *System*	*Driver*	*Driver*	Limited
2	**Partial Driving Automation**	The *sustained* and *ODD*-specific execution by a *driving automation system* of both the *lateral* and *longitudinal vehicle motion control* subtasks of the *DDT* with the expectation that the *driver* completes the *OEDR* subtask and *supervises* the *driving automation system*.	***System***	*Driver*	*Driver*	Limited
ADS* ("*System*") performs the entire *DDT* (while engaged*)			*System*	***System***	*Fallback-ready user (becomes the driver during fallback)*	Limited
3	**Conditional Driving Automation**	The *sustained* and *ODD*-specific performance by an *ADS* of the entire DDT with the expectation that the *DDT fallback-ready user* is *receptive* to *ADS*-issued *requests to intervene*, as well as to *DDT performance-relevant system failures* in other *vehicle* systems, and will respond appropriately.				
4	**High Driving Automation**	The *sustained* and *ODD*-specific performance by an *ADS* of the entire *DDT* and *DDT fallback* without any expectation that a *user* will respond to a *request to intervene*.	*System*	*System*	***System***	Limited
5	**Full Driving Automation**	The *sustained* and unconditional (i.e., not *ODD*-specific) performance by an *ADS* of the entire *DDT* and *DDT fallback* without any expectation that a *user* will respond to a *request to intervene*.	*System*	*System*	*System*	**Unlimited**

ADS: Automated Driving System, DDT: Dynamic Driving Task, OEDR: Object and event detection and response, ODD: Operational domain design.

1. 'Don't have secrets': HAL 9000 must not lie to humans, and must not distort or conceal information.

2. 'Be secretive': HAL 9000 must not tell really object of the space mission to human crew members.

These two instructions were staying in conflict with each other, and these generated a kind of cognitive dissonance for AI. HAL 9000 tried to reduce the 'stress' and selected a radical method for dissonance reduction: he began to murder the humans because they are

the central element of this cognitive dissonance. Namely, if there are no humans on board of spacecraft, then HAL 9000 doesn't have to hide anything from them, at the same time, this AI must not tell really object of the space mission to them.

Conclusions

The story of HAL 9000 can be analyzed in an interdisciplinary approach.

The psychological approach is: in the future when we have as AI as HAL 9000, psychologists will need to diagnose, prevent, and intervene in the cognitive (or non-cognitive) problems of AI (maybe therapy will need to be secured).

The educational approach is: likely there will be needed to teach people to give adequate instructions to robots (e.g. in order to avoid cognitive dissonance).

Moral philosophical and legal approaches are: although the solution for cognitive dissonance reduction of HAL 9000 may be the right from a logical viewpoint, this solution is unacceptable from aspects of morals, and law (especially human rights). This example raises the question that the strong (highly developed) artificial intelligence of the future will make decisions only on a logical basis or it will take into account moral and legal considerations as well. This question is especially current in such a world, what keeps straight to the futuristic autonomous vehicles, homes, and medicinal (Zorkóczy, 2021a,b), educational, business (Gulyás, 2022), etc. service providers (what will operate without human factor).

The informatical approach is: from the side of software making, it is needed error handling algorithms that can prevent or handle the expected or observed conflicts between instructions, and that can evaluate the effects and moral and legal aspects of the expected result of own operation. Last but not least: in the case of cognitive dissonance reduction, there is needed that this algorithm can select the best possible solutions, not only from logical, but moral and legal aspects too. In Isaac Asimov's robot novels, this security apparatus is not part of the software but part of the hardware. Maybe because Asimov thought that the hardware is less modifiable or vulnerable than software. The 'positron brain' of his robots contains the Three Laws of Robotics and these laws give safety to humans.

The Three Laws of Asimov's robots:

First Law: A robot may not injure a human being or, through inaction, allow a human being to come to harm.

Second Law: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.

Third Law: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law

Later Asimov added a zeroth law for his robot universe:

Zeroth Law: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Notes: Asimov's robots have more types, for example, softdroid, cardioid, android, humanoid droid, computer kind, animal-kind. Robots with a central brain control more bodies, etc. So, artificial psychology must adapt the new and new AI.

Frequently problems of Asimov's robot are (Mező, 2021):

1) people the robot is preferred instead of humans,

2) three laws' conflict (spontaneous or wanted),

3) the concept of 'human' is problematic.

Underlying the last conceptual problem are some additional difficult questions. These:

What is the prototype of a human?

How can an AI acknowledge that a person who is different the prototype of a human (e.g. a mentally and/or body-handicapped person) is human?

If humans are different from animals by the human mind, then a robot with strong AI is the same as a human or more (smarter, stronger, faster, more adaptable, etc.) than a human?

What should be more important for an AI: the individual human in a given moment or the humanity in the long run? Note: the 0th Laws of Asimov's robots prefer humanity instead of an individual, and the 0th law allows robots to kill if this is in the interest of humanity).

The answer to these questions depends only on human decisions - for now because AI can also have a say in this later. It's all just a matter of programming. The question is, of course, who or what writes the program... Who or what makes decisions? Who or what makes a programming error or issues malicious or misunderstandable commands?

If we assume an AI that can be programmed or directed through speech, will a malfunction similar to a psychological problem caused intentionally or accidentally be treated?

Does anyone know a good robot psychologist...?

Does anyone know Susan Calvin's phone number?*

References

AI HLEG (2019) *A definition of AI: Main capabilities and disciplines, High-Level Expert Group on Artificial Intelligence of the European Commission.* Downloded: 01.12.2022.URL: https://www.aepd.es/sites/default/files/2019-12/ai-definition.pdf

Adams, D. (1985): *The Hitchhiker's Guide to the Galaxy: The Original Radio Scripts.* London: Pan Books. 1985. ISBN 0-330-29288-9

Anderson, L.W. and Krathwohl, D.R. (eds.) (2001): *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman. ISBN 978-0-8013-1903-7

Asimov, Isaac (1993): *Robottörténetek 1-2.* Móra Ferenc Könyvkiadó, Budapest.

Bledsoe, W.W. (1966). Some Results on Multicategory Pattern Recognition. *J. ACM.* 13 (2): 304–316. doi: 10.1145/321328.321340

*Dr. Susan Calvin is a fictional robot psychologist character in Asimov's Robot Universe.

Bledsoe, W. W. and Chan, H. (1965): *A Man-Machine Facial Recognition System-Some Preliminary Results, Technical Report PRI 19*A, Palo Alto, California: Panoramic Research, Inc.

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H. and Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay Company

Cerasi, Isabelle Roche Cerasi (2016): *Automated Vehicles*. SINTEF Technology and Society Department of Safety and Mobility. Downloaded: 01.12.2022. URL: https://nvfnorden.org/wp-content/uploads/2020/10/06-Autonomous-vehicles-state-of-the-art.pdf

Clarke, Arthur C. (1968): *2001: A Space Odyssey*. London (UK): Hutchinson

Crowder, J. A. & Friess, Sh. (2012): *Artificial Psychology: The Psychology of AI*. Conference paper (March 2012)

Diagnostic and statistical manual of mental disorders: DSM-5. — 5th ed. (2013). W ashington, DC: American Psychiatric Association, ISBN 978-0-89042-554-1

Dick, Philip K. (1968): *Do Androids Dream of Electric Sheep?* New York (USA): Doubleday

Freeland, B. (2018): *Should We Be Worried About Cybernetic Mental Illness?* Downloaded: 22.06.2019. URL: https://medium.com/@benfreeland/should-we-be-worried-about-cybernetic-mental-illness-97a97460f604

Guilford, J.P. (1988). Some changes in the structure of intellect model. *Educational and Psychological Measurement*, 48, 1-4.

Gulyás Dávid (2022): Csevegőrobotok alkalmazása a customer service területén. *Mesterséges intelligencia – interdiszciplináris folyóirat, IV. évf. 2022/2. szám.* 27-41. doi: 10.35406/MI.2022.2.27

Hsu, F-h. (1999): IBM's Deep Blue chess grandmaster chips. *IEEE Micro* (Vol. 19, Issue: 2, Mar/Apr 1999), 70-81. doi: 10.1109/40.755469

Mező Ferenc (2021): Olvasókörök szerepe a mesterséges intelligenciával kapcsolatos attitűdök formálásában – Módszertani javaslat Asimov robot-történeteire reflektáló vitaklubok szervezésével kapcsolatban. *Mesterséges intelligencia – interdiszciplináris folyóirat*, III. évf. 2021/2. szám. 79-95. doi: 10.35406/MI.2021.2.79

Mező Ferenc és Mező Katalin (2019): Interdiszciplináris kapcsolódási lehetőségek a mesterséges intelligenciára irányuló cél-, eszköz- és hatásorientált kutatáshoz. *Mesterséges intelligencia – interdiszciplináris folyóirat*, I. évf. 2019/1. szám. 9–29. doi: 10.35406/MI.2019.1.9

Mező Ferenc, Mező Katalin és Mező Kristóf Szíriusz (2019): Filmklubok szerepe a mesterséges intelligenciával kapcsolatos attitűdök formálásában. Mesterséges intelligencia – interdiszciplináris folyóirat, I. évf. 2019/1. szám. 67–94. doi: 10.35406/MI.2019.1.67

Net1: Bloom's Taxonomy. Downloaded: 01.12.2022. URL: https://www.valamis.com/wp-content/uploads/2022/09/blooms-taxonomy-original.jpeg

SAE (2016): *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. Downloaded: 04 May

2020, Web: https://www.semanticscholar.org/paper/Taxonomy-and-Definitions-for-Terms-Related-to-for/5962a3287865a8453ddc7832340df322ea0f0bd0 doi: 10.4271/j3016_201609

Zorkóczy Miklós (2021a): AI Vision #1 – Goodbye asymmetric information for Patients?! *Mesterséges intelligencia – interdiszciplináris folyóirat*, III. évf. 2021/1. szám. 9-17. doi: 10.35406/MI.2021.1.9

Zorkóczy Miklós (2021b): AI Vision #2 – Medical Law and Artificial Intelligence in the Future of Healthcare. *Mesterséges intelligencia – interdiszciplináris folyóirat*, III. évf. 2021/2. szám. 9-18. doi: 10.35406/MI.2021.2.9

MÓDSZERTANI TANULMÁNYOK