Regular Article

Determinism and Social Paradoxes of Explainable Artificial Intelligence (XAI)

Soichiro Toda [1]

Eisuke Nakazawa [2]

Abstract

The core philosophical issue of explainable artificial intelligence (XAI) is the philosophical implication and clarifying the meaning of the term explanation. We call the process of connecting the cognitive suspension that exists between artificial intelligence (AI) and humans individually "explanation." If, through explanation, an AI is recognized as a moral agency, then and only then is the AI allowed to act in a way that satisfies the person as an XAI. AIs and humans have different operating systems to begin with. However, the condition that an AI is a moral actor, which equals a decision maker, is crucial for the AI to be recognized as an XAI. Furthermore, an XAI as a moral actor eliminates the paradox of infinite regress of explanations in the XAI argument. As an aid to this understanding, we examine the requirements for the social implementation of XAI, using the ethically interesting case of triage as a starting point. Then, we highlight the practical/philosophical paradox that cannot be resolved: can XAI create a story for explanation? We also discuss the trade-off between "accuracy" and "humanity" provide further topics for future research.

Keywords: explainable artificial intelligence, AI, philosophy, moral agency

1. Introduction

News of computers that use artificial intelligence (AI) technology, such as Alpha GO, defeating humans in games has been capturing much attention. However, the following considerations are also true: What if the AI suddenly played a Go stone in the corner in its first move? Questions such as, "Why did it make such an unorthodox move? There must be some secret, but we have no idea what it is...," would abound.

Nevertheless, Alpha GO could provide answers to those questions. For example, Alpha GO may show that this type of move is the most efficient way to beat the opponent by scoring the next predicted move. Certainly, the scoring must be within the range of human understanding to constitute an explanation. According to Hagras [1], the development of explainable artificial intelligence (XAI) progresses through three main methodologies: 1) deep learning; 2) interpretive models that incorporate causality; and 3) deductive methods that make black-box models immediately explainable.

The explanatory potential of AI contributes to the advancement of science. An example is the elucidation of protein structures in biochemistry and physical chemistry: the structure of a protein, which can be formed from

[1] Department of Philosophy, Tohoku University Graduate School of Arts and Letters, Corresponding Author (E-mail: soichirotoda@gmail.com)
[2] Department of Biomedical Ethics, University of Tokyo Faculty of Medicine

20 different amino acids, has been predicted and continuously updated using deep learning. However, when an evolutionary interpretation of that structure is required, the black-box nature of AI provides no answer; the AI merely presents a random catalog of protein structures. Undoubtedly, a better understanding of protein structure can be achieved through coevolutionary and interpretive explanations [2] in the same family of protein structures than through a mere catalog. Moreover, in 2022, DeepMind, an AI research company, announced that deep learning has revealed nearly all structure types that proteins can theoretically take (https://alphafold.ebi.ac.uk). A database of more than 200 million of these proteins has been made available to the public at no cost, and if the abovementioned explanations were added, research would accelerate further.

2. Who Is Explaining, to Whom, and How ?

There is a growing demand for XAI that can provide explanations. The term folk psychological XAI might also be applied here because it is the words and symbols used by humans that persuade humans. Thus, we use XAI to refer to technology that can be explained and understood in human language [3]. Importantly, such XAI also has significant implications for social theories of science and technology (e.g., [4]), and it will provide benefits as AI technology advances. Its advantages should be fully exploited in fields including medicine, disaster response, sports, agriculture, and bioscience.

As illustrated earlier, human lives may one day be in the hands of XAI, especially with the use of big data, which continues to attract attention. Although still in the trial stages, the application of XAI to medical big data and the elaboration of its procedures of clinical trials are remarkable. An example is the monitoring of the relationship between diagnosis, medication, and the patient's daily life in psychiatry. It has been reported that in psychiatry (especially in the diagnosis and subsequent treatment of schizophrenic patients), treatment according to situational judgment based on outpatient reports has limits [5].

Therefore, moderation theory is the general response to the “who does it, toward whom, and in what manner,” that is, “the appropriate XAI (and its analyst), to the appropriate person, in an appropriate manner.” However, the question as to who will make the clinical decision (especially end-of-life decisions) remains open. Furthermore, “why” generates another level of “because,” and that “because” generates another “because,” without even making philosophy of action. In this infinite regress, another very important problem system is where to break the chain of reasons in a humanly understandable way [6].

Therefore, let us consider two contexts with easy to analyze but ones we rarely pay attention to them. The former example is related to human emotion, and the latter is related to degrees of explanations.

The first is related to medical clinical practice. We will list typical diseases (but unrealistic in the actual medical practice), but please assume that the following description of the nature of the disease is the correct one derived by XAI.

> Ms. A has breast cancer and continues to receive outpatient care but is at a crossroads. According to medical AI, although mastectomy reduces the risk of death (a higher 5-year survival probability), profiling (AI analysis) of Ms. A’s constitution and genes shows that 7 or 8 years

after the resection, she is very likely to develop liver cancer (for an unknown reason). The probability of liver cancer is very high 7 to 8 years after the resection (for an unknown reason). Conversely, treatment with breast conservation has a lower 5-year survival probability than resection but a very low chance of liver cancer. XAI has made the decision that mastectomy is the best option because she must survive first.

How should the XAI be evaluated for making such a decision? The most appropriate explanation may be that the patient needs to survive first. Most patients would probably agree with this explanation. If XAI and AI propose the same rationale for mastectomy, what is the difference that XAI makes with respect to AI? It is, after all, “Ms. A’s narrative” that is necessary for an “adequate explanation,” even if the same conclusion is reached. Certainly, there is an upper limit to the amount of data that can be input for this task (implementation of Mr. A’s narrative into the AI), so it is important how much data is input. However, we must ask what is meant by the term “show of skill?”

Here, we would like to introduce a distinction between “heavy XAI” and “light XAI.” A heavy XAI is an AI based on a dataset of information that corresponds to the formation of Ms. A’s narrative by collecting as much information about Ms. A as possible. A light XAI, by contrast, is used when Ms. A's narrative is not necessary. The triage problem addressed in the next section will apply light XAI.

However, there are many situations in which a light XAI will deal with “heavy” decisions and the presentation of reasons for those decisions. The problem of triage is that a light XAI must make decisions from a small dataset that leads to life-and-death issues, which have been primarily addressed by bioethics. A light XAI, precisely because it is light, forces the reconsideration of the much-discussed catchphrase of bioethics at the time of its emergence. That is, with AI, we will "decide who dies and who lives.”

The second context is the explanation of a phenomenon of a familiar word, “burning.” Why do flames burn in the way they burn: in some places with a blue-like color, in some places with a slightly reddish color, in some places without water, sometimes in an orderly manner, and sometimes as a disordered flame? This “why” of the phenomenon of combustion is explained late in the education process, usually in high school (often, in the senior year in high school). Furthermore, we cannot explain the phenomenon of combustion with mathematical or chemical formulas. To explain it, an understanding of the concept of energy and of chemical equations of dozens of steps is necessary. If XAI tried to explain turbulent conflagration [7], many people would not care. Furthermore, it could be even said that no explanation is necessary, except for engineers, because society operates without an explanation. It might be said that XAI is the one that provides the necessary explanation for theoretical and practical parties in this way.

3. Should We Follow AI Triage Decisions?

In this section, we consider triage, which focuses on lifesaving situations, as the topic of discussion. Usually, triage refers to a series of medical actions in large-scale emergency medical care settings, such as in the event of a disaster, in which treatment priorities are determined by assigning a color code to each patient (e.g., black marks are placed on patients who are unlikely to

survive). In such a setting, what type of triage algorithm should be followed? In the first place, what is the philosophical and practical difference between light XAI and heavy XAI in triage? Is it merely “degrees of explanations?”

In Section 2, we proposed that it is the decision maker's narrative that determines the lightness of the XAI; in XAI triage, there is a short time for decision-making. It is important to note that the shortness of time to decide is not incompatible with the need for extensive and “deliberative” explanations. That is an algorithmic version of cognitive System 1 and System 2 we human have. If possible, as much of the patient's background as possible should be known; otherwise, the very fact that the patient's life was not saved because of a “hasty” decision made by the AI can be perceived as a defeat for the XAI by the disaster whether the decisions made are by light XAI or heavy XAI.

Even more troubling is the fact that most people will ignore the justification for “triage due to lack of information” as described above [8]. The autonomy of XAI (the power of XAI in this area will only grow stronger) is now eroding the proposition that "the final judgment of general responsibility and rational justification is usually reserved solely for humans [9]. Furthermore, XAI explanations contain sufficient content to raise the social question of who is really making the decisions? Humans should fill in the eroded (gapped) propositions with a variety of rationalizations (this very task is explanation), but how humans, not XAI, will fill in the gaps will be a challenge for the future. At the very least an algorithm in AI should be implemented that explains which information is necessary and sufficient for an explanation in a short temporal time window. Why did you mention information b instead of information a as the reason for the action? Therein lies the black-box problem of XAI.

4. Triage, Utilitarianism, and XAI as Moral Agent

At the root of triage is saving lives and differentiation based on utilitarianism. However, do medical practitioners trust medical AI or XAI enough to differentiate patients [10]? A trustworthy AI can save many lives on a utilitarian basis and must be mentioned with XAI [11]. AI agents (robots) are often considered personalities that can be trusted and are moral actors in some contexts [9]; chatbots are a good example of this [12]. If chatbots are accepted as moral agents, a type of intimacy must be assumed [13,14] that includes laughing at each other, being sad, and logically convincing each other. The chatbot's feelings are considered and predicted to establish a conversation with it. Moreover, it can be called a hypothesized person or a hypothesized moral agency. These relationships are referred to as a new theory of mind for human–AI [15].

However, it would be difficult to agree with the conclusion that the XAI is also a moral actor from the discussion of intimacy, which is accepted to a certain extent, and an epistemological gap exists [16]. Consider, for example, a human being who keeps kicking a robot that seems to be working with a certain purpose (usually walking). The robot is given a certain degree of assumed personhood because humans “feel sorry” for the “abused” robot. However, since it is not expected that the robot will proactively act or take revenge on the person who kicked it in a way different from violence to the real humans, it would not easy necessarily to say that the robot is a moral agent. (See the following article by

Reuters. https://www.reuters.com/video/watch/idRCV00696E.) It would also be informative, both in developmental psychology and in XAI research, to observe how babies or apes react when shown videos similar to this one. To further the argument, intimacy is what XAI should learn. A distinction needs to be made whether intimacy refers simply to a friendship or is more rooted in a sexual relationship [17,18]. If the epistemological gap can be filled, then the explanation of the XAI will be exactly the explanation we seek from the XAI.

The next exploration focuses on the black-box nature of XAI: if XAI explanations are pragmatically familiar to society, is it necessary to dismantle the black box? Conversely, is it necessary to clarify at what point in the chain of reasoning the XAI in question used “because” [19]? In the next section, we argue that the black box of XAI is by no means a Searle-style black box [20] but that to make the infinite regress of reasons unquestionable in a black box is to violate the autonomy required of XAI.

5. Autonomy from the Perspective of XAI: Various XAI and Conditions for Social Demand

The concept of black box (i.e., we cannot understand and check every function or algorithm in AI therefore there is no transparency about the processing for the put-puts) , can be likened to the fact that a computer can manipulated without the knowledge of how CPUs and semiconductors work; it is a function to describe a function. This is not to refute functionalism but to merely show that the “Chinese room” argument [20] is no longer keeping up with the development of XAI.

Autonomy is not a philosophical or ethical argument but rather, political one because it undermine autonomy itself, especially when it comes to human’s nature argument, but simply refers to the ability of XAI to produce out-puts in response to its surroundings; the outputs appear to be calculated from the computability domain in computer science.

In sum, It is worth noting that XAI provides explanations in different ways depending on the explanatory method or the rules applied to the input data [21,22].

We will attempt to position XAI based on these “limitations.” By saying “limitations”, we indicate the levels of input and following limited typical levels of outputs. Let us begin by applying XAI to a Chinese room in the Searle style. We do not follow this thought experiment from its foundation, but the unchanging assumption is that the people in the room do not understand Chinese at all. This is equivalent to someone who wants to use a spreadsheet but knows nothing about the basics of programming or computer “grammar.” The Chinese input into the room is output as Chinese (whatever the input/output is, the person in the room will not recognize it as Chinese) according to a vast manual. The person in the room becomes a Chinese speaker based on ignorance. Now, let us show with an example that the series of operations performed in the "traditional room" are powerless against the input of value-added sentences.

Does Mr. A weigh more than 60 kg? (Input, Chinese) (1)
Manual treatment of persons in the room (2)
Yes, Mr. A weighs 65 kg (output, Chinese) (3)

The exchange in (1) through (3) is a question of fact, and the number of steps in the inference is one. If this "one-shot" (input-processing-output once-only) factual question is a factual question, then the Chinese room

argument seems to be valid. The person in the room follows the Chinese rulebook and produces an output. However, what if further value-laden questions and discussion (inputs and outputus) follows?

Is Mr. A's weight appropriate for a 40-year old? (Input, Chinese) (1')

Manual treatment of persons in a room (2')

Yes, 40-year old Mr. A weighs the right amount (output and Chinese) (3')

Do you like Mr. A with the right weight? (Input) (4')

The conversation ends here because it is beyond the range of responses that can be output by the AI. (5')

Why do you like/dislike Mr. A? (Input) (6')

(the conversation also ends here)

As described above, since the Chinese room does not have preferences, it cannot answer value-added questions (i.e., value-laden questions or moral questions), such as "Do you like Mr. A?" More importantly, value-added decision-making requires the implementation of self.

If the XAI could answer subjective, respondent-specific questions in a retestable and reproducable manner, that would be a great progress for XAI research field. In the "Moral Turing Test" [23], XAI must implement a preference for someone (or something), and there are many questions that cannot be answered without a preference. However, that is the same as analyzing the science of someone's arbitrary intentions and empathic abilities to self and the consequences of their decisions [22,23].

We must also mention the relationship between autonomy and the Chinese language room. AI autonomy can be conceptually divided into two categories:

(5-1) Inherent in the AI (attributed to some person's self, including fictitious).

(5-2) Epistemological, as defined by the human observing the AI.

Note that (5-1) is intrinsic, such as preferences mentioned earlier, and autonomy (5-2) is defined by the human who observes it. XAI performs the decision content of decision-making (autonomy (5-1)), and for the assessment of epistemological autonomy, it is necessary to understand the process of decision-making (process of understanding) (5-2). Given that XAI is formed by a large number of modules, the process of understanding the process is dispersed and expressed at various levels.

However, two problems emerge as follows:

(a) The problem of formulating a theoretical coping policy for the "infinite regress of explanations of explanations of explanations of explanations..." between modules of XAI.

(b) The practical question of how and to what extent those who observe the process of breakdown of the decision-making by XAI must request a breakdown (the problem of arbitrary stops).

From these considerations and thought experiments, a conceptual framework for conducting the preliminary experiment aforementioned is needed. For example, the issue of reproducibility requires consideration: when looking at the response of XAI at times t1 and t2, other conditions being equal, the response obtained should be

the same. However, the paradox here is as follows: when humans reason and make decisions at times t1 and t2, the final responses may differ in the reasoning process. This is linked to "humanness," “fluctuations [24–26],” and “perturbations [27],” which we discuss below. However, if XAI does not make identical decisions at t1 and t2, that is, if reproducibility is not ensured, social implementation will be difficult. Imagine that an innocent suspect in a certain case is waiting for the sentence to be handed down. Then, along comes an "ambiguous (human-like)” XAI. As for a defendant, (S)he will think the sentence is not fair and unlucky. There is a possibility that the suspect’s life was determined by physical fluctuations (or perturbation). Still, the sentence is literally beautifully laid out and logical by a plausible explanation.

Again, the explanation in the Chinese room thought experiment is the primary role of the XAI accompanying the AI (the person in the room). Even if that XAI has a learning function, its ability to use its knowledge to make social decisions depends on the physical fluctuations of the XAI, which mimics human cognition. Knowledge and the representation of knowledge are not always in the mind of the individual [28]. The individual here is a human being, but this paper assumes that XAI also makes such representations of knowledge and predicts the associated effects on its surroundings. When value-added questions were asked, it was confirmed that it is the interrelationship between the multilayered consequences and reasons for each output that matters. However, it cannot be left to XAI with its probabilistic fluctuations to make decisions that affect a person’s life. However, have there not been attempts to implement XAI in such situations?

Moreover, these interrelationships run through the narratives that are input into the XAI. In Section 2 we introduced and distinguished between heavy and light XAI. Heavy XAI performed narrative exchange at the output–input layer and layered them as much as possible. Thus, difficulties appear when XAI makes its characteristic reason rise from those layers. The information obtained is very important, but its cost is enormous. Furthermore, the more complex the layered network in the implemented XAI, the more possibility—or contingency—is involved in social decisions. The more complex the network of layers in XAI, which requires social implementation, the more possibility in social decision-making, making social implementation difficult. This is the paradox in the social implementation of XAI.

6. Toward Social Implementation of XAI: Arbitrariness and Anthropocentrism

To address the paradox in the social implementation of XAI, we argue that it is useful to introduce arbitrariness and anthropocentrism into XAI. First, arbitrariness of XAI implies that

> according to the algorithm, the data is decoded to the point where it is human-interpretable, resulting in an “arbitrary suspension” of explanation on the part of the human.

Second, introducing anthropocentrism into XAI would modify the paradox in the social implementation of XAI. In this paper, XAI has been positioned initially as a so-called AI, an agent based on deep learning. Its algorithms, however, turn it into an agent that emphasizes a very human, qualia-like element that also makes

mistakes in its explanatory reasoning. For example, in explaining why music is played in a palliative care department, the agent (XAI) is human, and the way it explains is also recognized as human. Moreover, the present generation, when con-fronting the XAI, want such an explanation. However, if people ask themselves whether they would use or trust such XAI in practice, the answer is no, especially in causal explanations [29]. It would be agreeable to have an XAI that is human-like in its explanations, that provides a straight-forward answer, and that is accurate. However, there is a trade-off between humanity and accuracy as the goal of XAI.

This may lead to the question, “Does XAI have to be human?” From the discussion in Sections 1 and 2, we con-firm that XAI (in this paper) aims to answer difficult questions in everyday language. According to our position in this paper, in between the "human" agent and the "AI-like" (i.e., black box) agent is algorithmic bias, discrimination, and dogmatism; humans know that those are morally wrong. It is also an important mission to objectify discrimination by AI from a social psychological perspective and to evaluate the appropriate "distance" between humans and XAI in databases. How can such an appropriate distance be achieved? We cannot reconcile humanity and accuracy, as dis-cussed, if we (rightly) side with either side. However, as long as people remain human (and are forced to shoulder the human in a human-centric society), the object of XAI is to be human. The more it has its own story/plot, the more credibility (trustworthy or not) it will gain. Moreover, we are compelled to shoulder XAI with anthropocentrism (i.e., human-centric society) [30]. In other words, humans have an innate nature to give XAI a narrative (we henceforth call this the “story sufficiency theory”). There is no doubt that this is an issue worthy of consideration in the philosophy of XAI. In the story sufficiency theory, stories and plots that do not require explanation are also considered. For example, it may be concluded that in a large rectangle, a triangular figure appears to be chasing a round figure, and that, from a cognitive science point of view, even an infant would perceive a story/plot. A cognitive-philosophical perspective on chatbots may answer this question (see Mizukami [31] for details). We agree with Mizukami but believe that it is important to separate storytelling/plotting from moral agency: there will be occasions when it is necessary to think of chatbots or human-centric XAI in the context of storytelling/plotting without moral agency.

Worse, the story sufficiency theory makes the level of explanatory content of XAI and its prediction increasingly difficult. Stories/plots are unnecessary in medical practice, for example, where tumor detection is paramount. Conversely, in cognitive science, storytelling is the primary task of XAI because the decision-making process is central in the famous "Sally and Anne” experiment [32].

If the above assumptions are appropriate, there are three issues to be concerned about. First, XAI will have to be individualized (division of labor), and if tailor-made XAI [33] is not realized, the problem that underlies this pa-per—social implementation of XAI—is not plausible because the implementation of general (universal) XAI is virtually impossible. Second, the individualized division of labor will be difficult to achieve: it will require individually tailor-made profiling of the infinite number of profiles in XAI's addressees, and ultimately XAI will become "too heavy”. Accompanying

that, a problem of cost would appear. Finally, the ultimate goal of XAI is to help people (and patients) tell their own stories and use them in treatment and education, or in other words, to help them create narratives. Thus, the direction of the explanation is reversed. The following question must now be answered: "Can multiple moral agents in an XAI be simultaneously accountable for both events and cognitions?"

7. Conclusion

This paper takes the naivest position of XAI (a position that is surrounded by many exceptions but is still worth considering) and, after obtaining two paradoxes from the literature survey about its feasibility, discusses new findings about the direction of explanation. An XAI is arbitrary and objective at the same time (heavy/light XAI and subjective arbitrariness). It also simultaneously seeks two directions, from cognitive understanding to event understanding and from event understanding to cognitive understanding, depending on the state of the explainer (related to the story sufficiency theory). To resolve this paradox, or to say that it is not a paradox, we need professionals who are well-versed in fields such as computer mathematics, logic, and ethics, and, lay persons who intuitively confirm the reproducibility of XAI's decisions and judge that XAI is trustworthy. Therefore, open science and interdisciplinary research are required.

Acknowledgments:

We would like to thank Shogo Arai, Graduate School of Engineering, Tohoku University (now Faculty of Science and Technology, Tokyo University of Science), who provided useful advice in writing this paper. We also thank Takayuki Kira (Faculty of Law, Aichi University) and Tomohisa Sumida for their important suggestions. The authors would like to express their gratitude to all of them and to acknowledge that all responsibility for the text in this paper rests with the authors.

References

[1] Hagras, H. Toward human-understandable, explainable AI. *Comput.* 2018, *51*, 28-36. DOI: 10.1109/MC.2018.3620965.

[2] Liu, Y.; Palmedo, P.; Ye, Q.; Berger, B.; Peng, J. Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst.* 2018, *6*, 65-74.E3. DOI: 10.1016/j.cels.2017.11.014.

[3] Miller, T. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 2019, *267*, 1-38. DOI: 10.1016/j.artint.2018.07.007.

[4] Dhanorkar, S.; Wolf, C. T.; Qian, K.; Xu, A.; Popa, L.; Li, Y. Who needs to know what, when?: Broadening the Explainable AI (XAI) Design Space by Looking at Explanations Across the AI Lifecycle. In *Designing Interactive Systems Conference* 2021, 28 June-2 July 2021, ACM, pp. 1591-1602. DOI: 10.1145/3461778.3462131.

[5] Roessner, V.; Rothe, J.; Kohls, G.; Schomerus, G.; Ehrlich, S.; Beste, C. Taming the chaos? Using eXplainable Artificial Intelligence (XAI) to tackle the complexity in mental health research. *Eur. Child Adolesc. Psychiatry* 2021, *30*,

1143-1146. DOI: 10.1007/s00787-021-01836-0

[6] DiMarco, M. Wishful Intelligibility, Black Boxes, and Epidemiological Explanation. *Philos. Sci.* 2021, *88*, 824-834. DOI: 10.1086/715222.

[7] Niioka, T. Techno Life Selected Book "Burning": From Candles to Rocket Combustion, Ohmsha. 1994. (Japanese)

[8] Gold, A., Greenberg, B., Strous, R., & Asman, O. When do caregivers ignore the veil of ignorance? An empirical study on medical triage decision–making. *Medicine, Health Care and Philosophy* 2021, *24*(2), 213-225.

[9] Lötsch, J.; Kringel, D.; Ultsch, A. Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics* 2022, *2,* 1-17. DOI: 10.3390/biomedinformatics2010001.

[10] Wilkinson, D.; Zohny, H.; Kappes, A.; Sinnott-Armstrong, W.; Savulescu, J. Which factors should be included in triage? An online survey of the attitudes of the UK general public to pandemic triage dilemmas. *BMJ Open* 2020, *10*, e045593. DOI: 10.1136/bmjopen-2020-045593.

[11] Washington, P.; Yeung, S.; Percha, B.; Tatonetti, N.; Liphardt, J.; Wall, D. P. Achieving trustworthy biomedical data solutions. *Biocomputing 2021: Proceedings of the Pacific Symposium,* 2020, pp. 1-13. DOI: 10.1142/9789811232701_000 1.

[12] Nadarzynski, T.; Miles, O.; Cowie, A.; Ridge, D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digit. Health* 2019, *5*, 2055207619871 808. DOI: 10.1177/2055 207619871808.

[13] Brandtzaeg, P. B.; Skjuve, M.; Følstad, A. My AI Friend: How Users of a Social Chatbot Understand Their Human-AI Friendship. *Hum. Commun. Res.* 2022, *48,* 404-429; DOI: 10.1093/hcr/hqac008

[14] Lee, M.; Park, J. S. Do parasocial relationships and the quality of communication with AI shopping chatbots determine middle-aged female consumers' continuance usage intentions? *J. Consum. Behav.* 2022, *21*, 842-854. DOI: 10.1002/cb.2043.

[15] Wang, Q.; Saha, K.; Gregori, E.; Joyner, D.; Goel, A. Towards a mutual theory of mind in human-AI interaction: How language reflects what students perceive about a virtual teaching assistant. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* Yokohama, Japan, 8-13 May 2021, ACM, pp. 1-14. DOI: 10.1145/3411764.3445645.

[16] Jacquet, B.; Baratgin, J. Mind-reading chatbots: we are not there yet. In *International Conference on Human*

Interaction and Emerging Technologies. Paris, France, 27-29 August 2020, Springer, Cham, pp. 266-271. DOI: 10.1007/978-3-030-55307-4_40.

[17] Neveu, F. On the Philosophy of Mathematics: Reflections on "Making Science", Based on Cavaillès. In *Making Sense, Making Science*; Guillaume, A., Kurts-Wöste, L. Eds.; ISTE London, UK, 2020, pp. 45-62. DOI: 10.1002/9781119788461.ch4

[18] Vilone, G.; Longo, L. Classification of explainable artificial intelligence methods through their output formats. *Mach. Learn. Know. Extr.* 2021, *3*, 615-661. DOI: 10.3390/make3030032.

[19] Ohtsubo, N.; Nakae, T.; Fukazawa, Y. et al. XAI (Explainable AI): How did Artificial Intelligence Think Then? Ric Telecom Co. 2021

[20] Searle, J. R. Minds, brains, and programs. *Behav. Brain Sci.* 1980, *3*, 417-424. DOI: 10.1017/S0140525X00005756.

[21] Gerdes, A.; Øhrstrøm, P. Issues in robot ethics seen through the lens of a moral Turing test. *Journal of Information, Communication and Ethics in Society* 2015, *13(2)*, 98–109.

[22] Wallach, W.; Allen, C. Hard problems: framing the Chinese room in which a robot takes a moral Turing test. University of Birmingham, AISB/IACAP, *5*.

[23] Kim, H.; Byun, S. Designing and Applying a Moral Turing Test. *Adv. Sci. Technol. Eng. Syst. J.* 2021, *6*, 93-98. DOI: 10.25046/aj060212

[24] Mencar, C.; Alonso, J. M. Paving the way to explainable artificial intelligence with fuzzy modeling. In *International Workshop on Fuzzy Logic and Applications,* Genoa, Italy, 6-7 September 2018, Springer, Cham, pp. 215-227. DOI: 10.1007/978-3-030-12544-8_17.

[25] Chimatapu, R.; Hagras, H.; Kern, M.; Owusu, G. Hybrid deep learning type-2 fuzzy logic systems for explainable AI. In *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Glasgow, UK, 19-24 July 2020, IEEE, 2020, pp. 1-6. DOI: 10.1109/FUZZ48607.2020.9177817.

[26] Mendel, J. M.; Bonissone, P. P. Critical thinking about explainable AI (XAI) for rule-based fuzzy systems. *IEEE Trans. Fuzzy Syst.* 2021, *29*, 3579-3593. DOI: 10.1109/TFUZZ.2021.3079503.

[27] Galli, A.; Marrone, S.; Moscato, V.; Sansone, C. Reliability of explainable artificial intelligence in adversarial perturbation scenarios. In *International Conference on Pattern Recognition.* 10-15 January 2021, Springer, Cham, pp. 243-256. DOI: 10.1007/978-3-030-68796-0_18.

[28] Todayama, K. What I expect from social psychology from elsewhere: ..... In *Mind and Society in Science;* Karasawa, K., Todayama, K. Eds.; University of Tokyo Press, 2012. (Japanese)

[29] Schraagen, J. M.; Elsasser, P.; Fricke, H.; Hof, M.; Ragalmuto, F. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 2020 (Vol. 64, No. 1, pp. 339-343). Sage CA: Los Angeles, CA: SAGE Publications.

[30] Ehsan, U.; Wintersberger, P.; Liao, Q. V.; Mara, M.; Streit, M.; Wachter, S.; Riener, A.; Riedl, M. O. Operationalizing human-centered perspectives in explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan, 8-13 May 2021, ACM, Article no.94. DOI: 10.1109/MC.2018. 3620965.

[31] Mizukami, T. Possibilities and Limitations of the Concept of Virtual Actorhood in the Ethics of Social Robots" *Emerging Researchers Research Note* 2020, 27-36. (Japanese)

[32] Baron-Cohen, S.; Leslie, A. M.; Frith, U. Does the autistic child have a “theory of mind”? *Cognition* 1985, *21(1)*, 37-46.

[33] Schoonderwoerd, T. A.; Jorritsma, W.; Neerincx, M. A.; van den Bosch, K. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *Int. J. Hum. Comput. Stud.* 2021, *154*, 102684. DOI: 10.1016/j.ijhcs.2021.102684

Received 15 November 2022

Final version accepted 10 December 2022