sciendo

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 69 (82) 2024

DOI: 10.2478/slgr-2024-0024

Paweł Stacewicz
Warsaw University of Technology
Warsaw, Poland
e-mail: pawel.stacewicz@pw.edu.pl
ORCID: 0000-0003-2500-4086

Krzysztof Sołoducha
Military University of Technology
Warsaw, Poland
e-mail: krzysztof.soloducha@wat.edu.pl
ORCID: 0000-0003-1351-5487

THE TURING TEST AND THE ISSUE OF TRUST IN AI SYSTEMS[1]

Abstract. The Turing test, which is a verbal test of the indistinguishability of machine and human intelligence, is a historically important idea that has set a way of thinking about the AI (artificial intelligence) project that is still relevant today. According to it, the benchmark/blueprint for AI is human intelligence, and the key skill of AI should be its communicative proficiency – which includes explaining decisions made by the machine.

Passing the original Turing test by a machine does not guarantee that the machine is, from the point of view of a human, trustworthy; on the contrary, the idea of a machine capable of "fooling" or "outsmarting" a human is inherent within the concept of the test.

We postulate, that a test that guarantees a high level of trust should be: a) non-imitative, b) non-behavioral, c) focused on the ability to explain, taking into account, however, the design and rules of operation of the machine (and not just human expectations), d) focused on the machine's ability to learn.

Keywords: Artificial Intelligence, Turing test, trust, principle of limited trust, explaining, machine learning.

1. Introduction

The concept of the **Turing test**, originally called the *imitation game* [Turing, 1950], is a persistent and, to this day, highly inspiring element in the discussion of artificial intelligence (AI). Literally speaking, the concept refers to a certain method of testing computer systems for intelligence. It involves a free conversation between a human and a machine: if, in the course

of it, a competent judge cannot **distinguish** the system from a human, he must conclude that he is dealing with an intelligent entity, and with a more radical approach: with a thinking entity[2].

To put it a bit more broadly, the Turing test is not just a test, but a certain **historically significant idea** that has shaped the way we think about the AI project for many years. According to this idea, artificial intelligence should be built in the **likeness of** humans: it should therefore solve problems undertaken by humans (e.g., decision-making), perform human cognitive processes (e.g., logical inference), and present its results in a human-understandable form. An extremely important component of this way of thinking is the emphasis **on the communicative proficiency of machines**. According to the idea of the test, the ultimate measure of similarity to humans is the fact that an intelligent system is fluent in natural language and is engaged in meaningful linguistic interaction with humans.

Turing himself emphasized the importance of these skills not only as part of the test formula itself, but also as part of a **general reflection on the future of** AI **research**. Already in his 1948 research report, *Intelligent Machinery*, he identified three important areas of future applications that fall under the broad umbrella of communication. These are learning languages, translating from one language to another, and communicating with a machine in natural language [Turing, 1948, p. 2]. It must be acknowledged that today, in the era of learning, multimedia and highly interactive systems, these areas are at the core of AI research [Russel, Norvig, 2020].

The machines being developed today, of course, exhibit a much richer and more widely socially influential range of interactions with humans than the concept of the Turing test suggests. In addition to their ever-improving linguistic competence, today's AI systems genuinely take over some tasks from humans and make decisions that are vital to them. This raises important questions about **trust**[3]. Will a machine with intelligence comparable to that of a human, and therefore maximally versatile, evolving and autonomous, not pose a threat to humans? Will it operate according to our instructions and expectations? Won't it turn out to be clever enough to hide its real goals, posing only to pursue goals set by humans?

Elaborating on the idea of an intelligence test, Turing (1950) did not explicitly address the issue of trust. He did, however, write about human fears of the creation of intelligent machines that would, if they came into existence, undermine the culturally established vision of humans as privileged beings, "superior to the rest of creation" [Turing 1950, p. 444]. These fears, or some form of **ultimate mistrust** of hypothetical machines of the future, were, according to Turing, an important psychological factor[4] leading

his contemporaries to strongly believe that the construction of intelligent machines was fundamentally impossible. Let's add for the sake of argument that Turing himself argued with such a belief.

Regardless of such an extreme understanding of (lack of) trust, the question can be raised as to whether the original Turing **test** is a **test of intelligence that engenders trust** in a narrower sense? Will a machine that passes it be trustworthy to us in terms of the tasks assigned to it and the information provided? And if it won't, are there any valuable elements inherent in the idea of the test that allow us to formulate the boundary conditions of a good test for trust? These are the questions we will take up in the following sections of the paper.

2. Factors that strengthen trust in artificial intelligence systems

The relationship of human trust in information systems – as a relationship of a psychological nature, of which humans are the main party – must have its mental pattern in **interpersonal relationships** (see [Stacewicz, 2023])[5]. This observation goes hand in hand with the idea of the Turing test, which encapsulates the general postulate of maximum similarity between an AI system and a human. That the degree of fulfilment of this postulate is increasing is evidenced by gigantic advances in various areas of research and application[6]. Today's AI systems not only have great computing power but are able to use this power in a purposeful and increasingly versatile manner, thus inherent to human intelligence. They draw conclusions from the data they are provided with, make valuable hypotheses, provide ever better explanations for their decisions, and learn in various ways, increasingly going beyond reproductive and strictly programmed activities (see [Russel, Norvig 2020], [Davies et al. 2021]).

Progress in these exemplary areas is so significant that both in public discourse and in actual technical projects, there is no longer talk of intelligence per se, but of certain **higher-order features** that would, in addition to intelligence, condition and guide the decisions made by AI systems. They would therefore have similar functions to those of humans, whose intelligence is always embedded in a broader psychological, social and ethical context. This would include features such as acting in accordance with the law, the ability to provide credible explanations, or the ability to make complex ethical reasoning or filter the solutions provided due to an imposed set of ethical norms (see [Lara, Deckers, 2021], [Allen, Varner, Zinser 2000]). Such features are seen as crucial to the programme of developing trustworthy AI

(see [Ethics guidelines for trustworthy AI, 2019], [Polak, Krzanowski, 2023]). Although it should be added that, according to Moor's classical typology, machines have the possibility to reach the level of explicite moral subjects – without the chance of full subjectivity [Moor, 2006].

Focusing our attention on technical issues – related to the way the system works, rather than to certain non-IT criteria for choosing certain actions rather than others (such as ethical criteria) – we can identify three factors that significantly strengthen users' trust in the system. These are:

- **effectiveness** (understood more broadly as robustness) – the system effectively solves the problems falling within its scope of competence (the better the statistics of correctly performed tasks, the greater the user's confidence);
- **explainability** – the system is able to explain, in a convincing way for a human, why in a given situation it has made a certain decision and not another (one possible way of explaining is a human-understandable reconstruction of the steps that led to the decision)[7];
- **flexibility and openness to criticism** – the system can change the way it makes decisions depending on the interaction with the user; especially in the case of ineffective actions, wrong decisions, etc...

Related to the last two factors is what we could call the *principle of limited trust*: the user **does not trust the system unlimitedly**, taking into account the possibility that it may make mistakes. In the light of this principle, trust in the system is gradual, and the more the system is able to eliminate errors and improve its operation, the higher the level of trust must be. This is facilitated by the **learning modules**[8] present in the system's software, but also, in addition, by the ability to generate explanations, which is strongly linked to corrective.

We pose the thesis that with the acceptance of the principle of limited trust, the role of the **ability to generate explanations** as a factor strengthening the reliability of the system increases significantly. If the user accepts the principle, and is therefore aware of possible errors in the operation of the machine, it naturally links credibility with the ability to improve this operation. This involves a long-term perspective that goes beyond the narrow horizon of the tasks currently being performed. The greater the **corrective capabilities** of a system, the higher the level of trust in it in the long-term perspective of current and future tasks.

The mechanism for generating explanations is extremely important in this context. By generating explanations, the system, firstly, satisfies a certain epistemic need of the user – the need to know justifications or explanations[9]. Secondly, however, it provides the user with a strong reason to believe that there is a **monitoring mechanism**

in the system, which, in addition to generating explanations, allows relevant errors to be automatically recognised and removed. In addition, the information contained in the explanations, especially those relating to the internal parameters of the system, can help to **improve the system** by a human, e.g. a programmer. In the context of a possible redesign, such a system is therefore more trustworthy than a system that does not provide explanations.

While we have identified both related factors – the ability to generate explanations and the ability to learn – as having a positive impact on trust, it is important to note another, more **negative** aspect of their interplay. The learning mechanism, or more precisely the changes that arise as a result of its operation, is considered to be one of the main causes of the **epistemic opacity** of computing systems and the difficult interpretability of the results they produce [Zednik 2021], [Stacewicz, Greif, 2021]. The typical process of increasing opacity is that, as a result of learning, certain technical parameters are introduced into the system to adapt the operation of the system to the requirements of the task being performed, but they do not allow to grasp a meaningful and human-understandable relationship between data and results. This problem is exacerbated when the learning process depends significantly on random steps, which (again) are intended to provide efficiency to the system rather than a human-transparent form of explanation. Thus, it turns out that although the ability to learn results in an improvement in the quality of the system's performance and, as such, strengthens the trust relationship, in the case of certain forms of learning – namely those that lead to epistemic opacity of the system – the trust relationship is **weakened**.

A possible solution to this problem is to link learning to explanation in such a way that **the same mechanism** for monitoring system performance is involved in both processes. On the one hand, the results of this mechanism, i.e. the logged changes of the system, would be the basis for learning (beneficial changes in the system) and, on the other hand, they would be used in the explanation process (which also refers to the internal states of the system). We will return to this issue in the last chapter, postulating the necessary conditions for a good trust test.

3. Is the Turing test a good test for trustworthy AI?

The original Turing test is aimed at testing whether a machine, guided by appropriate software, manifests intelligence **sufficiently similar to human intelligence** in conversation. Recall: if a competent judge, conversing

with a machine and a human on any subject, cannot distinguish between them, he must consider the machine to be intelligent [Turing, 1950].

It is therefore an **imitation** test, as Turing himself emphasised when he called it an *imitation game*. Crucially, however, the outcome of the test depends on the ability to imitate purely external manifestations of intelligence, i.e. machine-generated utterances. The internal and technical aspect of the machine is ignored[10].

With regard to the actual capabilities of a machine, resulting from its internal design, this should be understood in such a way that it must, at least in certain situations, **conceal its full capabilities**. This means that if the system under test differs from a human in certain important respects – e.g., the speed of operation, the way data is processed or the precision of the results generated – it should still imitate human behaviour in order to pass the test. Sometimes, therefore, in order to pretend to be human, a machine will give wrong, incomplete or slow answers[11]. And if its program gives the possibility of reaching correct and fast answers, then, de facto, in such situations, the machine hides its actual capabilities.

Note that the strategy of maximizing similarity to humans also applies to **explanations**. In order to pass the test, the machine must provide the kind of explanations that the judge expects – the kind of explanations, therefore, that a human would provide, rather than the kind of explanations that are consistent with its way of operating.

Can a machine imitating a human being in a total way – pretending to be one and hiding its real capabilities – engender trust? Certainly not. From a psychological point of view, such traits as a tendency to pretend (someone or something else) or to **conceal** one's own intentions and capabilities are not conducive to building trust. On the contrary, they must give rise to suspicions that the other party in the trust relationship is not really who he or she claims to be and may behave in unexpected ways, including **unsafe** ones.

In a less psychological context, it can be argued that the original Turing test does not provide the human (judge) with a complete knowledge of the machine, giving him an insight into a very **sparse slice of its capabilities** (or more precisely: their external manifestations). This lack of knowledge must undermine trust. Referring, in turn, to the internal knowledge of the machine, the knowledge on the basis of which the machine operates, we must conclude that the Turing judge has no access to the essential components of this knowledge.

From an even broader point of view, abstracting from the specifics of the Turing test in general, it has to be said that intelligence itself, tested in

any case, does not necessarily **go hand in hand with credibility**. Indeed, intelligence can be used for various purposes and in a variety of ways. Thus, if an artefact passes the intelligence test, it cannot automatically be considered credible or trustworthy. This is because it is not impossible that it will use its intellect in a way that is not beneficial or even threatening to us.

To conclude this thread, it must be said that the verbal test of human-machine indistinguishability proposed by Turing **is not a good test for trustworthy AI**. In the original conception of the test, there is an irreducible contradiction between answering the judge's questions fairly, taking into account the real capabilities of the machine, and being required to give answers (including: explanations) that are indistinguishable from human ones. Nonetheless, there are some valuable elements in both Turing's original proposal and its critique that may bring us closer to a test formula that takes into account trust in AI.

4. Proposal: boundary conditions for a good test for trustworthy AI

Although the original Turing test is not a good test for a trustworthy AI, the idea behind it – the idea of machines **communicating** with humans in natural language – points, in our opinion, in the right direction. Valuable and effective communication, i.e. communication that is based on understanding and seeks to understand the issues at stake, must include explanation; and we have identified the latter as a factor that is significantly conducive to trust. Referring back to Turing's own proposal [1950], it must be said that the ability of machines **to explain** their decisions was indeed something he pointed to as something we should verify in a test. This is evidenced by the examples he cites of conversations in which questions are asked to explain why one solution was chosen and not another[12].

Keeping the main idea of the Turing test, but also bearing in mind its various shortcomings, we can propose four **boundary** conditions for a good test of trustworthy AI (abbreviated as TTW test). Their juxtaposition will serve as a kind of summary of the theses formulated above.

Firstly, emphasising once again our main thesis, an essential element of the whole procedure should be the checking of the machine's ability to **give explanations**; although this element should be integrated into a more general idea, which the following points bring closer.

Second, a good TTW **should not be a purely imitative test**. The machine should not so much mimic the verbal behavior of a human, but

model itself after them, keeping its own **difference and identity**. In particular, if a piece of the test conversation would be focused on explanation, a competent judge should pay attention more to the consistency of the explanations and their conformity to the specifics of the task given to the machine than to their similarity to the explanations given by humans. With such an approach, the tester would not depreciate original and creative answers (those differing from typical human explanations), on the contrary: he would appreciate them.

Third, a good TTW test **should not be a behavioral test** that ignores the internal rules of the AI system. In particular, the explanations generated by the system should be checked for compliance with these rules – rules that the judge knows (as opposed to the judge in the original Turing test). For example: the system could justify its decision by citing a particular algorithm implemented in it and the data processing technique behind it. If the decision turned out to be wrong, the system would not completely lose trust, since it pointed to an algorithm that could be improved in the future thanks to this indication.

Fourth, a good TTW test should target the **ability to learn**, which in the long run, by being able to correct erroneous or suboptimal actions of the system, is a factor that strengthens trust[13]. In checking this ability, the tester would not only ask questions, but would provide the system with some knowledge or present it with solutions to certain problems, and then check to what extent the system is able to change its way of operating to a more effective one.

We have called the proposed conditions **boundary** conditions, because they apply only to certain aspects of the system's operation – functional aspects, related most strongly to information technology and the way AI systems work (e.g., whether they are capable of learning, whether they are equipped with explanatory functions, etc...). These conditions, however, do not apply to certain non-IT criteria governing the choice of goals pursued by the system and the way they are pursued, including legal and ethical criteria (see EU document on trustworthy AI). For example: such criteria could exclude actions aimed at harming people (physical or psychological) or spreading untruth.

Although questions of compliance with the law and ethical norms were not the main concern of the paper, it is worth clarifying at the end that a special variant of the verbal indistinguishability test called the moral Turing test is considered in analyses devoted to the ethics of AI. It involves testing the degree of similarity between artificial systems and competent moral agents [Allen, Smit and Wallach, 2005; Allen, Varner and Zinser, 2000;

Moor, 2006; Krzanowski, Trombik, 2021]. As with the original TT, the testing procedure is in the form of an interview, focused, however, on ethical evaluations and decisions. That is, the judge asks questions like "What would you do in a situation involving ethical decisions regarding someone's health or life?" or "Can you recognize whether a certain ethical decision, set in a so-and-so described context, was made by a machine or by a human?". Based on the participant's answers and broader explanations, the judge makes a verdict on who is a machine and who is a human, and on this basis the machine (unless it is recognized as a machine) can be considered a moral agent. An additional type of tests are also ethical reasoning tests in which the system is supposed to make a moral conclusion based on the information gathered and the moral patterns it has (Arnold and Scheutz, 2016).

Although the moral Turing test targets ethical issues relevant to the trust relationship, by virtue of the same arguments we have posed for the standard TT (see section 3), it is not a good test for trustworthy artificial intelligence. To be such, it would have to meet the boundary conditions we postulated. Thus, it should be **non-imitative**, **non-behavioral**, and aimed at testing **learning** and **explanatory** abilities. In fact, the evaluation of the correctness of a moral agent's behavior concerns to a large extent not only the effect of an action itself, but often the evaluation of efforts leading to the avoidance of potential bad consequences of some activity. And in such situations, insights into a machine's decision-making process are important for assessing its moral competence.

NOTES

[1] The research presented in this article was supported by National Science Centre (NCN) OPUS 19 grant no. 2020/37/B/HS1/01809.

[2] Such radical interpretation is hinted at by Turing himself, who, in his 1950 text, directly formulates the question "Can a machine think?".

[3] In this text we understand the concept of trust broadly, going beyond the domain of interpersonal relationships towards (human – artificial system) relationships. In the field of artificial intelligence analysis, this is a common practice, as evidenced by the commonly used term "trustworthy AI".

[4] Of course, this was not a factor of a logical nature, because from fear of machines or mistrust of them in no way follows the thesis that it is impossible to create machines with intelligence comparable to that of humans.

[5] In the referenced work, I delve deeper into the psychological aspects of trust relationship. I adopt the definition, that the trust is a psychological relationship, which consists in the fact that a certain person, let's call him *a subject of trust*, has a certain kind of belief in the possible actions of another party, for example, a person, animal, institution, device or system. Namely, this person is convinced that these actions will be in accordance with his expectations, which, in turn, have their justification in the declarations,

commitments or technical specifications (in the case of artifacts) of the other party. This way of understanding trust is also adopted in this work.

[6] It is worth noting that the postulate of the similarity of the system to a human has received a strong development in the conception of expert systems. Although these systems are narrowly specialised, their essence is that a particular system is to show equal effectiveness in a given field as a human expert. It is therefore supposed to be indistinguishable from a human expert. See [Gupta, Nagpal 2020].

[7] Concrete systems vary greatly in terms of the interpretability of the results they generate and, going further, the ability to explain the decisions they make in terms that are understandable, i.e. easily-interpretable, by the user. This depends primarily on the method of knowledge representation used in the systems. In the case of symbolic representations, such as decision trees or semantic networks, the ability to explain is high. In the case of sub-symbolic representations, including connectionist representations used in artificial neural networks, it is low, and generating meaningfully transparent explanations requires the use of sophisticated algorithms whose reliability is still debated. See [Stacewicz, Greif, 2021], [Zednik 2019].

[8] Learning modules are already standard in modern AI systems. Moreover, the ability to learn is increasingly being identified as a defining feature of artificial intelligence. See [Woolridge, 2021].

[9] This need is inherent to the nature of man as a rational being i.e. striving to obtain knowledge. Every epistemological definition of knowledge emphasises the factor of justification: knowledge is a set of beliefs which, among other properties, must have a sufficiently good justification (see [Ajdukiewicz, 2006], [Chisholm 1977]). The latter, in turn, in a form accessible to humans i.e. intersubjectively communicable, takes the form of an explanation that answers the question 'why?'. For example: 'Why did the doctor or the AI system replacing him/her make a particular diagnosis and recommend a particular therapy?'.

[10] It is worth noting that this weakness of the test was emphasised by John Searle when he formulated his famous Chinese Room argument [Searle, 2010]. According to his critique, a machine operating correctly with Chinese expressions, that is, generating utterances indistinguishable from those of the Chinese, may not understand their meaning at all, and this, according to Searle, obviously negates its intelligence. However, we, as persons assessing the machine for intelligence, can recognise this lack of understanding (gain certainty in this respect) only if we delve into the way the machine works (not its external manifestations). A test that ignores the internal rules of operation therefore does not give true information about the similarity of the machine's intelligence to human intelligence (which is founded on understanding). See [Stacewicz 2022].

[11] In fact, this is how Turing puts the matter, discussing in his paper short examples of questions and answers [Turing, 1950].

[12] Here is a few-sentence example in which the interrogator asks why the witness (we can think of it as the equivalent of a machine) suggested such and not a different beginning of the sonnet:

«Interrogator: In the first line of your sonnet which reads 'Shall I compare thee to a summer's day', would not 'a spring day' do as well or better?
Witness: It wouldn't scan.
Interrogator: How about 'a winter's day'. That would scan all right.
Witness: Yes, but nobody wants to be compared to a winter's day.»
[Turing 1950, p. 446].

[13] Note that Turing himself emphasized the strong connection between learning ability and intelligence. In the same paper in which he formulated the rules of the machine intelligence test, he devoted an entire subsection to hypothetical machine learning techniques;

further pointing out that only machines capable of learning would be able to pass the test [Turing 1950].

BIBLIOGRAPHY

Allen C., Smit I., Wallach W., (2005), *Artificial morality: top-down, bottom-up, and hybrid approaches*, "Ethics and information technology", volume 7, s. 149–155.

Allen C., Varner G., Zinser J., (2000), *Prolegomena to any future artificial moral agent*, "Journal of Experimental & Theoretical Artificial Intelligence", Volume 12, 2000 – Issue 3, 251–261. DOI: 10.1080/09528130050111428.

Ajdukiewicz K., (2006), *Zagadnienie uzasadniania*, [w:] Język i poznanie, t. 2, s. 374–384, Warszawa, Wydawnictwo Naukowe PWN.

Arnold T., Scheutz M., (2016), *Against the moral Turing test: accountable design and the moral reasoning of autonomous systems*, "Ethics and Information Technology", 18, s. 103–115. doi.org/10.1007/s10676-016-9389-x.

Chisholm R.M., (1977), *Theory of Knowledge*, Prentice-Hall inc.

Davies A., Veličković P., Buesing L. et al., (2021), *Advancing mathematics by guiding human intuition with AI*, "Nature", 600, 70–74 (2021). https://doi.org/10.1038/s41586-021-04086-x

Drozdek A., (1998), *Human Intelligence and Turing Test*, "AI & SOCIETY", 12, 315–321.

Ejdys J., (2017), *Determinanty zaufania do technologii*, "Przegląd organizacji", 12/2017, 20–27.

EU High-Level Expert Group on AI, (2020), *Ethics guidelines for trustworthy AI*, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Gerdens A., Øhrstrøm P., (2015), *Issues in robot ethics seen through the lens of a moral Turing Test*, "Journal of Information, Communication and Ethics in Society", 13(2), 98–109. DOI: 10.1108/JICES-09-2014-0038.

Gupta I., Nagpal G., (2020), *Artificial Intelligence and Expert Systems*, Mercury Learning and Information.

Krzanowski R., Trombik K., (2021), *Ethical Machine Safety Test,* Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?, Cognitive Technologies, Springer Nature Switzerland AG 2021. https://doi.org/10.1007/978-3-030-56546-6_10.

Lara F., Deckers J., (2020), *Artificial Intelligence as a Socratic Assistant for Moral Enhancement*, "Neuroethics" 13, 275–287 (2020). https://doi.org/10.1007/s12152-019-09401-y

Mishra M., (1996), *Organizational responses to crisis: The centrality of trust*, [in:] R. M. Kramer, T. Tyler (ed.) "Trust in Organizations", Newbury Park, Sagae, p. 261–287.

Moor J. H., (2006), *The nature, importance, and difficulty of machine ethics*, "IEEE Intelligent Systems", 21(4), 18–21.

Polak P., Krzanowski R., (2023), *How to Tame Artificial Intelligence. A Symbiotic AI model for Beneficial AI*, Ethos, 3(143), 92–106.

Russell S., Norvig P., (2020), *Artificial Intelligence: A Modern Approach*, Pearsons, 4th ed.

Searle J.R., (2010), *Minds, brains, and programs*, Cambridge University Press.

Stacewicz P., (2022), Wyjaśnianie, zaufanie i test Turinga, [in:] M. Jakubiak, P. Stacewicz (ed.) "Zaufanie do systemów sztucznej inteligencji", Oficyna Wydawnicza Politechniki Warszawskiej, p. 23–34.

Stacewicz P., Greif H., (2021), *Concepts as decision functions. The issue of epistemic opacity of conceptual representations in artificial computing systems*, "Procedia Computer Science", 192, p. 4120–4127.

Turing A.M., (1950), *Computing Machinery and Intelligence*, Mind, New Series, 59 (236), p. 433–460.

Turing A.M., (1948), *Intelligent machinery: A report by A.M. Turing*, National Physical Laboratory, London.

Véliz C., (2021), Moral zombies: why algorithms are not moral agents, "AI & SOCIETY" 36, 487–497. DOI: 10.1007/s00146-021-01189-x.

Woolridge M., (2021), *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*, Flatiron Books.

Zednik C., (2019), *Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence,* "Philosophy & Technology", 2019, https://doi.org/10.1007/s13347-019-00382-7.