Open Access Library Journal
2023, Volume 10, e10870
ISSN Online: 2333-9721
ISSN Print: 2333-9705

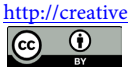
Advancing Trustworthy Explainable Artificial Intelligence: Principles, Goals, and Strategies

Zeeshan Sadiq, Muhammad Aqib

College of Computer Science and Technology, Taiyuan University of Technology, Taiyuan, China
Email: shadi0045@link.tyut.edu.cn, jixiang0038@link.tyut.edu.cn

How to cite this paper: Sadiq, Z. and Aqib, M. (2023) Advancing Trustworthy Explainable Artificial Intelligence: Principles, Goals, and Strategies. *Open Access Library Journal*, **10**: e10870.
https://doi.org/10.4236/oalib.1110870

Received: October 10, 2023
Accepted: November 27, 2023
Published: November 30, 2023



Abstract

In the contemporary era, artificial intelligence (AI) has introduced transformative advancements that have significant implications for society. Nevertheless, these advancements come with challenges, notably those associated with opacity, vulnerability, and interpretability. The integration of artificial intelligence (AI) systems into various aspects of human life has become increasingly pervasive. Consequently, there is a growing need to prioritize the development of trustworthy and explainable artificial intelligence (XAI) as a paramount concern within the field. The main purpose of this paper is to explore the paramount importance of XAI, clarify its multifaceted meanings, and outline which consists of a series of guiding principles essential for the development of XAI. These principles simultaneously act as overarching objectives, directing the course towards ensuring transparency, accountability, and reliability in AI systems. Additionally, the paper presents two novel strategies to actualize XAI, by narrowing the difference between AI's potential and human understanding. By addressing the intricate issues associated with XAI; this study adds to the continuing dialogue on how one might tap into the complete potential of AI technology, ensuring its responsible and ethical implementation in an ever-evolving digital environment.

Subject Areas

AI

Keywords

Explainable Artificial Intelligence, Advancement, Trustworthiness, Transparency, Accessibility

1. Introduction

Artificial intelligence technology, while conferring advantages to various fields

including science, the economy, and everyday life as well as showcasing its capabilities encounters certain challenges. For example, contemporary artificial intelligence deep learning models are evolving to be more substantial and intricate [1]. A model with an augmented number of parameters and heightened complexity often achieves enhanced accuracy, but this comes at the cost of its interpretability. Consequently, machine learning models are becoming increasingly opaque, resembling a black box. Moreover, these models exhibit a lack of resilience against adversarial attacks, a deficiency that introduces grave security threats in areas like autonomous driving. Gary Marcus and Ernest Davis have pinpointed nine challenges associated with present-day AI, which include fundamental attribution errors, a dearth of robustness, machine learning's profound dependency on the meticulous specifics of expansive training sets, and an indiscriminate over-reliance on data, which can perpetuate archaic societal biases, among other concerns [2]. Conversely, the pervasive implementation of artificial intelligence brings to the fore issues related to discrimination, bias, and privacy. Presenting a myriad of challenges to societal ethics and jurisprudence [3]. Notably, their potential utilization in pivotal and delicate sectors such as health care, finance, and autonomous lethal weaponry, which touch upon national well-being and security have attracted considerable scrutiny from global governments, culminating in the suggestion of diverse regulatory principles and goals [4].

Whether it is the issues of trustworthiness, robustness, or the black box (opacity) that current artificial intelligence research and development focus on, all are closely related to the explainability of artificial intelligence [5]. It can be stated that explainability is the prerequisite and foundation for addressing robustness, subsequently, and trustworthiness. Therefore, the topic of explainability artificial intelligence, commonly known as Explainable Artificial Intelligence (XAI), has become a central concern in the field of artificial intelligence.

Based on the argumentation of why the main point of explainable artificial intelligence (AI) is important, the paper precisely defines explainable AI. It elucidates the core concerns or objectives that explainable AI aims to achieve and further proposes some conceptual approaches to realize explainable AI.

2. Explain-Ability Is a Core Concern of Artificial Intelligence

Decisions made by artificial intelligence programs are typically derived from algorithms, which inherently convey an aura of objectivity. However, the exact processes by which these models or algorithms produce their results remain elusive. Developers were uncertain about how it determined the winning move. This uncertainty exemplifies the black box problem stemming from algorithmic opacity. Another manifestation of the black box issue occurs when developers, due to commercial confidentiality or other reasons, withhold training data. This secrecy can lead to ethical and social challenges, including bias and discrimination. Considering this, researchers like Crawford and colleagues in a research report advocated for a ban on the use of black box artificial intelligence and al-

gorithmic systems in core public institutions (e.g., those overseeing criminal justice, healthcare, welfare, and education) recommending a shift to systems that ensure accountability through methods such as validation, auditing or public scrutiny. They argue that understanding a system's weaknesses is the initial step towards its enhancement; thus, explanation becomes a vital component of future AI model training and validation processes [6]. Consequently, from the developmental perspective of AI explainability holds paramount significance.

2.1. Explainable Artificial Intelligence Is the Foundation of Trustworthiness

In the research literature concerning the explainability of artificial intelligence, a frequently cited example pertains to image recognition and classification. In this instance, a recognition model that integrates a deep neural network with logistic regression can accurately classify most images. However, it erroneously identifies a husky in the snow as a wolf. Further investigation into the model's explainability revealed that the classifier, based on its training data, associated large areas of snow with the identification of wolves. This correlation arose because, in the training samples, wolves consistently appeared against snowy backgrounds, whereas huskies did not. In experiments, human evaluators upon grasping such a decision-making rationale, diminished their trust in the model to 11% [1]. In a similar vein, a parking sign heavily adorned with stickers was mistakenly labeled by Google's automatic labeling system as a refrigerator filled with food and drinks. Such non-robust and unreliable artificial intelligence systems struggle to earn user's trust. Individuals will hesitate to utilize the associated products, especially AI products tied to user safety, like autonomous driving. Thus, trust in the development, configuration, and utilization of AI systems is not merely an intrinsic technical attribute and necessity but also a hallmark of the techno-social framework of AI applications. As Brian Christian observed, with the swift evolution of machine learning models in global decision-making systems, many individuals realize they possess limited insight into the internal workings of these models, leading to prevailing unease.

The issue of explainable artificial intelligence is that fundamentally pertains to human-machine interactions. Its importance can be underscored by recognizing the role of explanation in cultivating trust among individuals. Consider education as an illustrative example. In a particular secondary school, a teacher during a review of exam questions, endeavored to persuade students of a standard answer by offering an explanation. The objective was to encourage students to embrace and have faith in this answer, though some remained skeptical. However, in a subsequent exam, the identical question resurfaced, but the provided standard answer differed markedly from the earlier one. As a result, students expressed their reservations. The teacher in response, presented what appeared to be a cogent explanation for this new answer. After witnessing such inconsistencies on multiple occasions, several students began to harbor significant distrust toward the teacher.

The trust that the scientific community and cognitive entities broadly invest in scientific hypotheses is anchored in explanation [7]. Typically, the proposition and acceptance of a scientific hypothesis or theory correlate with its capacity to elucidate phenomena. For example, in contemporary consciousness science, which bears a close relation to genuine general artificial intelligence, mainstream theories gain acceptance largely because they can clarify certain pivotal phenomena and questions of public interest. The Higher-order theories, for example, concentrate on delineating the reasons underlying the consciousness of mental states. They adeptly clarify why certain contents are conscious while others are not, due to their inability to be the subject of appropriate meta-representational states. Meanwhile, the Global workspace theories can distinctly expound conscious access, signifying their capability to shed light on why specific representations can be adaptively employed by various cognitive systems functioning as information consumers [8]. In situations where multiple competing hypotheses are present, a hypothesis ought to surpass its rivals by offering a more comprehensive explanation of the target phenomena. It should exhibit enhanced explanatory powers, implying its capacity to coherently elucidate a more extensive spectrum of data or phenomena compared to its competitors. Additionally, it should seamlessly align with successful theories in related domains. Beyond possessing robust explanatory power, the philosopher of science, Lipton, posits that the best explanation should also be the loveliest explanation, denoting an explanation imbued with greater potential understanding [9]. Naturally, in the methodology of science, other factors influence the trust and acceptance of a hypothesis, including consistency both internal and in relation to background theories, simplicity, analogy, and among others [10].

The trust of the human cognitive community in science is rooted in explanation. In a similar vein, since artificial intelligence technology stems from human cognitive activity, the trust that humans place in artificial intelligence and its products should also be anchored in explanation. Generally, individuals approach technologies with caution if they cannot be readily explained are untraceable or seem untrustworthy [11]. From a social psychology standpoint, humans feel more aligned with and assured about entities they are familiar with, namely those whose operational mechanisms, methods, and outcomes they comprehend. Artificial intelligence systems and products that offer explainable are thus comprehensible to individuals and are more adapted to secure user's trust.

2.2. Explainability Is the Premise and Foundation of Artificial Intelligence Governance

The widespread adoption of artificial intelligence technology in daily life introduces potential risks and challenges in areas like ethics, privacy and law. Concurrently, its potential applications in high-risk sectors such as finance, healthcare and autonomous lethal weapons have attracted considerable attention from relevant governmental bodies globally. These positions AI governance as a cru-

cial element in the development and deployment of artificial intelligence.

When governmental regulatory authorities consider allow the use of artificial intelligence systems in society, they must evaluate several critical aspects, including compliance, safety, controllability, algorithmic transparency, privacy, data governance, robustness, and accountability. These evaluations are closely related to the explanation of artificial intelligence. For example, safety can be divided into two dimensions: the safety of the artificial intelligence system itself and the safety of humans. Regarding the safety of the artificial intelligence system, it should clearly indicate its vulnerable areas to attacks and potential forms of attack, such as data tampering, network breaches, infrastructure vulnerabilities, etc. How does the system respond in unforeseen circumstances or settings? What contingency measures does the system employ in the face of adversarial assaults? Regarding human safety, the system should provide a transparent assessment of the potential risks and magnitude of harm it could inflict on users or third parties. Additionally, it should detail the losses or detrimental effects that might ensue if the artificial intelligence system produces incorrect results or decisions.

Transparency is crucial in the governance of artificial intelligence algorithms. An efficient AI system should not operate as a black box but instead should offer clear explanations of how its algorithms make decisions. For example, in the banking sector, online loan applicants should understand why their loan applications were declined or why they were granted lower loan amounts than they requested. To achieve this, the system must articulate the methodologies used in the design and development of the algorithmic framework, as well as the techniques for testing and validating the algorithmic system. This includes the scenarios and cases used for testing and validation, as well as pertinent details about the data used for these tasks. Moreover, the AI system should explain the comprehensibility of its decisions and how they may impact an organization's decision-making process.

Contemporary artificial intelligence systems predominantly utilize deep learning technology, which has a significant reliance on data. The gathering and selection of data for training, testing, and validation are intrinsically tied to concerns like personal privacy and algorithmic fairness. Evaluated artificial intelligence systems ought to elucidate the sources, varieties, and extent of the data they employ. For instance, they should specify whether and to what degree they incorporate sensitive personal data and whether protective measures such as encryption, anonymization, and aggregation are implemented to safeguard personal privacy. It is also essential to determine if there exists discrimination and bias related to attributes like race, gender, wealth, education, and social status in the data collection process, as these biases could precipitate algorithmic (model) unfairness.

The practical application of AI faces a major obstacle in the form of explainability. It is difficult to understand why machine learning algorithms function in

specific ways and explain them to others. This is mainly because research and development departments are disconnected from the commercial market, creating a gap between technology and society. However, it is crucial to ensure transparency and trust in artificial intelligence systems. These systems must clearly state their objectives, capabilities, and limitations, and their decisions should be interpretable to the relevant users and administrators to an appropriate degree.

2.3. The Explainable Artificial Intelligence

Explainable artificial intelligence refers to artificial intelligence that is comprehensible. It is a characteristic of artificial intelligence models. As contemporary artificial intelligence predominantly relies on machine learning, explainability primarily pertains to the comprehensibility of machine learning models. It is the ability of a model to render its operational mechanisms clear to its audience.

In the research literature, the concept of explainability is not clearly defined and consistent. Furthermore, related concepts such as interpretability, transparency, and comprehensibility are often used interchangeably. Among these, comprehensibility is the most fundamental and is essentially linked with the other concepts. Comprehensibility refers to a model's ability to be understood without the need for an explanation of its internal structure or the algorithms used for data processing. It measures how effectively individuals can comprehend the decisions made by the model. Explainability is an active property of a model that includes any actions or procedures the model takes to clarify or refine its internal operations. Societally, explainability can be seen as the model's ability to ensure fairness. Interpretability, on the other hand, is a passive property of a model, indicating that it is an attribute that can be grasped or understood by human observers. The ability to offer explanations or meanings that are easy to understand is called transparency. This is important because it helps ensure fairness in decision-making by identifying and correcting biases in datasets. It can also improve the robustness of models by highlighting potential disruptions. Transparency also helps to establish the significance of variables, such as identifying genuine causal relationships in model inference. Comprehensibility refers to the ability of machine learning algorithms to present their knowledge in a way that humans can understand. It is important that learning algorithms can communicate their results in natural language so that people can interpret them. Comprehensibility is often related to the complexity of the model. The audience's ability to understand the information in the model is essential for comprehensibility, which is closely tied to interpretability.

The issue of transparency is intimately tied to the widely criticized black box dilemma in artificial intelligence. A model is deemed transparent if it is inherently understandable. The challenge of algorithmic transparency is fundamentally a socio-technical matter, as aspects such as the sources and methods of training data collection, the decision-makers involved, and their underlying motivations play pivotal roles in determining transparency. Given that various

models possess distinct motivations and degrees of comprehensibility, and these vary relative to different audiences like developers, users and managers, models can be classified based on their transparency levels. Adrian Weller has identified eight distinct levels of transparency, contingent upon different audiences and motivations. For instance, for developers, model transparency entails comprehending the workings of their systems, troubleshooting methodologies, and enhancement strategies. For users, the system should elucidate its actions and rationale, anticipate its behavior in unpredictable scenarios, and foster trust regarding the technology. Transparency and simplicity in gauging the influence of algorithmic input features on decisions not only empower users to grasp the determinations made by AI algorithms and their underlying reasons but also facilitate governmental regulatory entities and organizations in effectively overseeing the decision system's operations, ensuring regulatory adherence, and thereby fostering trust in the artificial intelligence system.

The issue of explainable artificial intelligence (XAI) is a matter of human-computer interaction. As the audience seeks explanations, human agents play a crucial role in elucidating the explainability of artificial intelligence. For example, D. Gunning defines explainable artificial intelligence as a suite of machine learning techniques that humans can comprehend, trust appropriately, and manage effectively. Aleatha *et al.* emphasize the importance of the audience, stating that explainable AI is the type of AI that can produce details or reasons to make its operation transparent or comprehensible to the audience. This definition highlights the human-computer interaction aspect but provides a general description of the content to be explained without explicitly delineating it.

Considering research on the topic of explainable artificial intelligence, we define explainable artificial intelligence as follows: relative to various human groups acting as the audience seeking explanations, it pertains to artificial intelligence that functions in a manner transparent and comprehensible to the specific audience, enabling them to grasp its learning, decision-making and predictive processes. Moreover, it should earn the audience's trust and adhere to regulatory standards. Within this framework, the human community acting as the explanation-seeking audience can be segmented into three categories: artificial intelligence system developers, regulatory officials in the domain of artificial intelligence, and users of artificial intelligence products. Different groups possess distinct expectations for explainability, resulting in varied levels of explainability. For example, for system developers, attributes like the security, robustness, and transferability of artificial intelligence systems hold paramount importance. Conversely, regulatory officials emphasize elements such as privacy awareness, protection, fairness, and accountability. Users of artificial intelligence products primarily value ease of access and clarity in decision-making processes, whereas the exactness and technical depth of explanations are of lesser concern. Explanations should be articulated in straightforward natural language and employ visualization or other readily comprehensible techniques.

3. The Primary Principles of Developing Explainable Artificial Intelligence

In line with the provided definition, interpretable artificial intelligence systems can be described as interactive frameworks grounded in human-computer interaction. Owing to differences in motivation, application contexts, tasks, users and other variables, the goals and prerequisites for the interpretability of artificial intelligence systems differ. However, some regulatory entities and researchers have proposed foundational principles for the creation of interpretable artificial intelligence. For example, the National Institute of Standards and Technology (NIST) in the United States has outlined four guiding principles to which interpretable artificial intelligence systems should conform: 1) the Explanation Principle, which dictates that the system must offer evidence or rationale for all its outputs; 2) the Meaningfulness Principle, which necessitates that the system provides explanations understandable to individual users; 3) the Accuracy of Explanation Principle, which ensures that explanations accurately represent the system's output generation process and 4) the Knowledge Limitation Principle, which maintains that the system should function only within its designed parameters or when it has ample confidence in its outputs. It's important to highlight that not all researchers support these four principles. For example, some like Watcher and colleagues, contend that the demand for precise explanations is overly rigorous, suggesting that a counterfactual explanation is adequate [12]. Chinese scholars have set forth more detailed criteria concerning the competencies that interpretable artificial intelligence systems ought to exhibit [6]. 1) intelligent agents should be capable of introspection and self-argumentation; 2) intelligent agents should demonstrate cognitive abilities and adaptability towards humans; and 3) intelligent agents should possess the capacity to create models.

It is readily apparent that the principles advanced by the regulatory entities and expert scholars predominantly focus on technical facets. The principles outlined by NIST have a broader scope, whereas those suggested by scholars are more detailed. The primary shortcoming of both principal sets is their neglect of the human aspect and their omission to situate the topic of interpretable artificial intelligence within the more expansive socio-technical framework we underscore. To rectify this, we introduce the subsequent four key principles for the evolution of interpretable artificial intelligence. These principles can also be viewed as essential criteria for interpretable artificial intelligence.

3.1. At the Core of This Concern Lies the Autonomy and Well-Being of Individuals

Artificial intelligence is a product of human development, and its main purpose is that AI is a remarkable creation from human ingenuity with the primary purpose of improving human well-being. In designing AI systems, it is essential to prioritize human-centeredness. The goal is to enhance our cognitive, social, and cultural abilities, reduce workloads, increase work efficiency, promote mental

and physical health, enrich cultural and entertainment experiences, and cultivate refined aesthetics. When operating AI systems, human control is critical. In case an AI system behaves abnormally or poses unforeseen risks, individuals must have the power to intervene, adjust, rectify, or shut down the system. Stuart Russell, a renowned AI expert, emphasizes the importance of designing machines that will defer to humans, seek permission, proceed cautiously when directives are ambiguous, and allow themselves to be deactivated. Purpose is to enhance human well-being, rather than create products that could threaten, manipulate, or subjugate people. When designing AI systems, it is crucial to adhere to the principle of human-centeredness. The objective should be to improve and complement human cognitive, social, and cultural abilities, reduce workloads, enhance work efficiency, promote mental and physical health, enrich cultural and entertainment experiences, and foster refined aesthetic sensibilities.

During the operation of AI systems, maintaining human oversight and control is crucial. If an AI system behaves abnormally or poses potential unforeseen economic and societal risks due to outputs beyond its design intentions, individuals must have the ability to intervene, adjust, rectify, or shut down the system. As the renowned artificial intelligence expert, Stuart Russell, points out, machines designed in this manner will defer to humans: they will seek permission, proceed cautiously when directives are ambiguous, and permit themselves to be deactivated. [13].

Interpretable artificial intelligence systems ought to advance social well-being and sustainable development. While these systems are crafted to achieve particular application tasks, tackle societal challenges and facilitate social progress, environmental considerations should remain paramount. Their development should not detrimentally impact the environment and ecosystems. In modern AI systems, the quest for accuracy frequently requires the use of vast amounts of training data and the adjustment of an increasing number of parameters. This leads to larger models and elevated energy consumption. Such practices can have negative environmental consequences, which may subsequently affect societal sustainability.

Data-driven artificial intelligence systems, when applied across various facets of social life, influence not only individual's social skills but also their interpersonal relationships. While these systems aim to enhance social competencies and diversify interactions, they can also bear consequences for individual's physical and mental health. For example, electronic communication via screens lacks the intimacy of direct interactions, making it difficult to foster emotional connections between individuals. This impersonal exchange of information might lead to a rise in individuals experiencing social anxiety disorder. Moreover, the design, implementation and use of interpretable artificial intelligence systems should prioritize fairness in interactions between individuals and societal entities, ensuring that these systems do not perpetuate bias or discrimination.

3.2. Reliability

The explainability of artificial intelligence is intrinsically tied to the trustworthiness of AI systems. As Russell noted, there is a consensus that if AI systems are to be trusted, their decisions must be explainable [14]. Artificial intelligence systems, being human-developed products designed to address specific human needs, necessitate broad societal acceptance and trust for widespread adoption. Thus, trustworthiness is a fundamental attribute of artificial intelligence and by extension, of interpretable artificial intelligence. Concurrently, evaluating the explainability of AI systems is deeply connected to the degree of trust humans place in the system and the system's reliability [15].

Trustworthiness is rooted in social and psychological foundations. Profound social interactions among human's hinge on trust in the participating parties, primarily influenced by two pivotal aspects: safety and consistency. People are more inclined to trust entities they perceive as safe. For instance, compared to formidable animals like tigers, humans favor smaller, less threatening creatures like cats and dogs, perceiving them as safer and less harmful. Moreover, individuals tend to trust those they are familiar with because they expect consistent behavior in similar situations, allowing them to anticipate the outcomes of their actions.

Artificial intelligence systems are at their core products utilized by humans. They serve as subjects of human communication and interaction. Consequently, interpretable artificial intelligence systems must also adhere to the two primary tenets of human social interactions: safety and robustness.

As previously noted, the safety requirements for interpretable artificial intelligence can be categorized into two facets: the safety of the artificial intelligence system itself and its safety in relation to individuals. Interpretable artificial intelligence systems should elucidate how they address adversarial attacks from sources like the internet and how they maintain robustness in diverse environments. Concurrently, they should clarify how they minimize the potential risks of causing harm to users or third parties in unexpected situations.

Another essential criterion for the trustworthiness of artificial intelligence systems is technical robustness. This includes both the capability of the artificial intelligence control system to sustain its functionality under exceptional conditions and the robustness of its models. The latter pertains to the model's capacity to uphold its performance and accuracy in the real world, beyond the controlled laboratory setting in which it was developed. A robust model should consistently produce correct judgments, predictions and decisions under similar inputs and contexts. This means its outputs should be repeatable and such repeatability should be adequately explained.

3.3. Transparency

Transparency is vital for the effective configuration and application of artificial intelligence systems in the real world [16]. It can be viewed as a foundational

requirement for interpretable artificial intelligence systems. On one hand, transparency is essential for explaining model decisions and overseeing artificial intelligence. On the other hand, it provides the foundation for evaluating interpretable artificial intelligence systems concerning privacy protection, fairness, and other aspects. Transparency covers elements such as data, models and business models.

The transparency of a model pertains to its intrinsic capacity to be comprehended by model developers, regular users, and relevant department managers, elucidating how it formulates judgments, decisions, or predictions. Transparency encompasses three distinct levels: algorithmic transparency, decomposability, and simulatability. These three are interrelated. For example, a simulatable model is also decomposable and algorithmically transparent. Simulatability pertains to the model's ability to be simulated or deeply understood by individuals, making complexity a pivotal factor. Generally, sparse linear models are more simulatable than dense linear models [17]. Decomposability refers to the model's capability to elucidate the entire model by interpreting its components, including inputs, parameters, and operations. Among the myriad AI algorithmic decision systems, only algorithms such as linear regression, decision trees, Bayesian methods, and k-nearest neighbors inherently exhibit transparency. When an algorithm itself lacks interpretability, its decomposability permits the construction of post hoc interpretable models based on the algorithm's internal structure, shedding light on the decision mechanisms and processes of the original decision system. Consequently, this characteristic can bolster the capacity to comprehend and elucidate model behavior. For a model to be algorithmically transparent, it must satisfy specific constraints, wherein each segment of the model should be comprehensible without the necessity for auxiliary tools. Algorithmic transparency can be perceived in various manners. It concerns the ability of users to grasp how a model processes inputs to produce a particular output.

It should be noted that in the context of artificial intelligence systems, greater transparency is not always preferable. Transparency can sometimes conflict with other attributes of artificial intelligence systems, as in certain scenarios, increased transparency might result in reduced efficiency, diminished fairness, and compromised trustworthiness [18].

3.4. Accessibility of Explanation

Contrary to the three principles previously mentioned, the accessibility principle relates to the form requirements of the explanations provided by artificial intelligence systems. This principle suggests that the explanations given by AI systems should be presented in a manner easily comprehensible to users and regulators affected by the model's decisions. Ideally, an interpretable model should alleviate any challenges in understanding that non-technical or non-expert users might face when interacting with such algorithms. In a review article on interpretable artificial intelligence, the author, after thorough examination, concludes

that this concept (*i.e.*, accessibility) is recognized as a third essential objective in the considered literature [19].

Accessibility primarily pertains to the accessibility of artificial intelligence systems for their general users. This means it allows individuals without expertise in the relevant technology to understand the evidence and reasoning upon which the system is based and how it reaches a particular decision. The accessibility of interpretable systems is intimately tied to the model's complexity and the sophistication of the techniques and tools used in its development. Generally, the fewer parameters a model has, the smaller and simpler it becomes, facilitating easier comprehension by individuals. Conversely, if the technical methods employed in constructing the model are intricate and challenging for non-technical individuals to understand, the accessibility of the artificial intelligence system using that model diminishes.

Accessibility necessitates that the explanations provided by artificial intelligence systems be both accurate and clear. Regarding content, explanations should include essential information about the system, especially its purpose, scope, functions, operating mechanisms, and details about the issues it addresses. Explanations should present this information in a scientifically accurate manner to prevent user misconceptions. In terms of presentation, if text serves as the medium for explanations, the language should be straightforward, using natural language and avoiding technical terms unless necessary. Visual representation, often referred to as visual explanations, is a beneficial option when feasible. Visual explanations can dynamically depict the model's behavior, highlighting its processes. Visualizations might also integrate additional techniques to elucidate intricate interactions among model parameters for users to improve clarity.

4. Conclusions

In conclusion, attaining interpretable artificial intelligence (AI) is a multifaceted challenge that necessitates addressing the inherent limitations of contemporary AI systems. These limitations, frequently manifested as the black box problem and concerns related to robustness, stem from the present state of AI, which lacks comprehensive knowledge and understanding. One AI expert notes that our machines do not effectively acquire, accumulate, apply, transmit and manage knowledge.

The journey towards achieving interpretable AI in human-computer interactions requires a profound exploration of the intrinsic structure of human cognition. This includes understanding the overarching principles that govern human cognitive processes, reasoning, and decision-making. In this context, two specific approaches emerge as essential:

Firstly, it is essential to construct AI systems endowed with human-like common sense and background knowledge, allowing them to understand the real world deeply. While AI systems based on deep learning might occasionally mi-

sinterpret everyday objects, such as mistaking a bus for a refrigerator, humans effortlessly make such distinctions due to their inherent common sense and contextual understanding. The specific knowledge and perception required might differ among AI systems designed for various purposes, but a foundation of common sense and context remains fundamental.

Secondly, there is an urgent need to integrate best-explanation reasoning with deep learning. Over the past seven decades, the AI community has debated whether symbolic reasoning or neural networks should form the foundation of AI systems. Currently, data-driven deep learning technologies are predominant. However, there is a growing consensus that these two approaches can coexist and complement each other. The challenge is in effectively merging them, often termed the Holy Grail problem in AI. This recognizes the potential of deep learning to include symbolic reasoning.

An optimal approach to this integration involves the development of a dual-layer AI system that combines best-explanation reasoning with deep learning. Deep learning continues to be the mature, mainstream technology, while best-explanation reasoning aims to uncover causal relationships within data and offer intuitive explanations. This reasoning model draws upon human background knowledge and common sense, with the knowledge produced acting as valuable data for deep learning models.

In summary, achieving interpretable AI requires a holistic approach that combines human-like common sense, background knowledge and best-explanation reasoning with deep learning capabilities. By integrating these elements, we can strive for AI systems that are not only powerful but also interpretable and aligned with human understanding.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Samek, W., Wiegand, T. and Müller, K.R. (2017) Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models.

[2] Knittel, B., Coile, A., Zou, A., Saxena, S., Brenzel, L., Orobaton, N. and Kasungami, D. (2022) Critical Barriers to Sustainable Capacity Strengthening in Global Health: A Systems Perspective on Development Assistance. *Gates Open Research*, **6**, 116. https://doi.org/10.12688/gatesopenres.13632.1

[3] Kim, P.T. and Bodie, M.T. (2021) Artificial Intelligence and the Challenges of Workplace Discrimination and Privacy. *Journal of Labor and Employment Law*, **35**, 289-315.

[4] Ribot, J.C., Agrawal, A. and Larson, A.M. (2006) Recentralizing While Decentralizing: How National Governments Reappropriate Forest Resources. *World Development*, **34**, 1864-1886. https://doi.org/10.1016/j.worlddev.2005.11.020

[5] Koshiyama, A., Kazim, E. and Treleaven, P. (2022) Algorithm Auditing: Managing the Legal, Ethical, and Technological Risks of Artificial Intelligence, Machine Learning, and Associated Algorithms. *Computer*, **55**, 40-50.

https://doi.org/10.1109/MC.2021.3067225

[6] Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019) Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, **9**, e1312. https://doi.org/10.1002/widm.1312

[7] Cosmides, L. and Tooby, J. (1994) Beyond Intuition and Instinct Blindness: Toward an Evolutionarily Rigorous Cognitive Science. *Cognition*, **50**, 41-77. https://doi.org/10.1016/0010-0277(94)90020-5

[8] Franks, B., Bangerter, A. and Bauer, M.W. (2013) Conspiracy Theories as Quasi-Religious Mentality: An Integrated Account from Cognitive Science, Social Representations Theory, and Frame Theory. *Frontiers in Psychology*, **4**, Article No. 424. https://doi.org/10.3389/fpsyg.2013.00424

[9] Campos, D.G. (2011) On the Distinction between Peirce's Abduction and Lipton's Inference to the Best Explanation. *Synthese*, **180**, 419-442. https://doi.org/10.1007/s11229-009-9709-3

[10] Simon, D., Snow, C.J. and Read, S.J. (2004) The Redux of Cognitive Consistency Theories: Evidence Judgments by Constraint Satisfaction. *Journal of Personality and Social Psychology*, **86**, 814-837. https://doi.org/10.1037/0022-3514.86.6.814

[11] Surden, H. (2019) Artificial Intelligence and Law: An Overview. *Georgia State University Law Review*, **35**, 19-22.

[12] Ramesh, A.N., Kambhampati, C., Monson, J.R. and Drew, P.J. (2004) Artificial Intelligence in Medicine. *Annals of the Royal College of Surgeons of England*, **86**, 334-338.

[13] Russell, S.J. (2010) Artificial Intelligence a Modern Approach. Pearson Education, Inc., London.

[14] Jiang, Y., Li, X., Luo, H., Yin, S. and Kaynak, O. (2022) Quo Vadis Artificial Intelligence? *Discover Artificial Intelligence*, **2**, Article No. 4. https://doi.org/10.1007/s44163-022-00022-8

[15] Brunette, E.S., Flemmer, R.C. and Flemmer, C.L. (2009) A Review of Artificial Intelligence. 2009 4*th International Conference on Autonomous Robots and Agents*, Wellington, 10-12 February 2009, 385-392. https://doi.org/10.1109/ICARA.2000.4804025

[16] He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X. and Zhang, K. (2019) The Practical Implementation of Artificial Intelligence Technologies in Medicine. *Nature Medicine*, **25**, 30-36. https://doi.org/10.1038/s41591-018-0307-0

[17] Oke, S.A. (2008) A Literature Review on Artificial Intelligence. *International Journal of Information and Management Sciences*, **19**, 535-570.

[18] Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S. and Zhang, J. (2021) Artificial Intelligence: A Powerful Paradigm for Scientific Research. *The Innovation*, **2**, Article ID: 100179. https://doi.org/10.1016/j.xinn.2021.100179

[19] Korteling, J.H., van de Boer-Visschedijk, G.C., Blankendaal, R.A., Boonekamp, R.C. and Eikelboom, A.R. (2021) Human versus Artificial Intelligence. *Frontiers in Artificial Intelligence*, **4**, Article ID: 622364. https://doi.org/10.3389/frai.2021.622364