Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Research paper

A manifesto on explainability for artificial intelligence in medicine

Carlo Combi [a,*], Beatrice Amico [a], Riccardo Bellazzi [b], Andreas Holzinger [c], Jason H. Moore [d], Marinka Zitnik [e], John H. Holmes [f]

[a] *University of Verona, Verona, Italy*
[b] *University of Pavia, Pavia, Italy*
[c] *Medical University Graz, Graz, Austria*
[d] *Cedars-Sinai Medical Center, West Hollywood, CA, USA*
[e] *Harvard Medical School and Broad Institute of MIT & Harvard, MA, USA*
[f] *University of Pennsylvania Perelman School of Medicine Philadelphia, PA, USA*

ARTICLE INFO

Keywords:
Artificial intelligence
Explainability
Explainable artificial intelligence
Interpretability
Interpretable artificial intelligence

ABSTRACT

The rapid increase of interest in, and use of, artificial intelligence (AI) in computer applications has raised a parallel concern about its ability (or lack thereof) to provide understandable, or *explainable*, output to users. This concern is especially legitimate in biomedical contexts, where patient safety is of paramount importance. This position paper brings together seven researchers working in the field with different roles and perspectives, to explore in depth the concept of explainable AI, or *XAI*, offering a functional definition and conceptual framework or model that can be used when considering XAI. This is followed by a series of desiderata for attaining explainability in AI, each of which touches upon a key domain in biomedicine.

1. Introduction

There is considerable discussion in the biomedical informatics and computer science communities about the “un-explainable” nature of artificial intelligence (AI), in that much is made of so-called “black-box” algorithms and systems that leave users, and even developers, in the dark as to how results were obtained. As a result, there is growing skepticism about the potential limits of AI, even in the face of burgeoning interest that at times reflects over-optimism about it. At the same time, there is a growing community of researchers who are working to address this skepticism through their work in making AI explainable, and thus useful and potentially usable to those who employ AI in their work. This is especially welcome in the domain of biomedicine, where explainable AI is critically important for clinicians in their daily practice.

As AI (including Machine Learning) becomes increasingly ubiquitous, there are growing concerns and questions, such as:

- How does an AI algorithm work — what is it doing?
- Does an AI system work as well as an expert?
- Does an AI system do what a user would do, were she in the same situation?
- Why cannot the system tell a user how it arrived at a conclusion or made a decision?

These concerns are of urgent importance and need to be addressed with scientific and engineering rigor in a variety of biomedical domains, including clinical decision support systems, patient monitoring, public health surveillance, and biomedical research. However, we in the informatics community are uniquely positioned to take leadership roles in developing and implementing strategies for improving the explainability of AI systems.

The primary goal of this paper is to present a compelling case for the need to address gaps in the explainability of AI software and the results presented to users. We hope to meet this goal by means of a rigorously developed conceptual model for thinking about explainable AI, or *XAI*, through a thorough exposition of the work to date and identification of gaps in research and application of XAI, and a proposition for how these gaps could be addressed. Even though many definitions and concepts we will introduce and discuss are general and may be applicable to many different domains, in the following we will focus on XAI in Medicine and Health. Indeed, these domains have special requirements that make XAI quite idiosyncratic and worthy of particular attention.

We have structured this paper as follows: after an introduction to the problem of explainability, in Section 2 we discuss some background on how informatics and computer science describe the problem, approaches to explainability, and applications of XAI to a variety of key clinical domains; Section 3 contains a proposal for a conceptual framework and foundational definition of XAI; Section 4 presents a set

[*] Corresponding author.
E-mail address: carlo.combi@univr.it (C. Combi).

https://doi.org/10.1016/j.artmed.2022.102423
Received 2 March 2022; Received in revised form 4 October 2022; Accepted 4 October 2022
Available online 9 October 2022

of desiderata that would be important to address XAI moving forward; finally, Section 5 sketches some conclusions and future directions.

2. Background

In this section, we will briefly introduce the main aspects that have been discussed about XAI in general, in the areas of Computer Science and Artificial Intelligence. Then, we will move to the main specific issues of XAI in Medicine, ending with some non-exhaustive examples of XAI approaches in clinical domains.

2.1. A research field's description of the current landscape of AI

The concept of explainability has a long story in AI. Indeed, since the first proposals of the so-called "expert systems", there was the need of having an explanation of why and how some conclusions were reached by the system in a complex decision-support task. Such a requirement was, and remains, extremely important in medicine, as physicians needed to understand why the system was proposing, for example, a specific diagnosis or treatment regimen. The need of having some explanation about the output, an AI-based system provides, has recently been exacerbated by the adoption of machine learning (ML) approaches, where the reasoning task is often performed by "black-box" systems that do not allow one to understand clearly why a specific result has been reached [1].

In principle, explainability is related to understanding, i.e., having a mental model of what we are observing. With a slightly different terminology, we may say that explaining/interpreting consists of providing causes of observed phenomena in a comprehensible manner through a linguistic description of its logical and causal relationships [1,2]. In the context of XAI, we need to understand the conclusions of a system that is reasoning on some data to reach some result. Such systems in medicine are often related to a decision-support task, where data may be incomplete, uncertain, ambiguous, or missing. Moreover, such data have a high complexity and heterogeneity, being expressed as often interrelated and intertwined data in various formats such as structured, semi-structured, or unstructured alphanumeric data, movies, images, sounds, waveform signals, and so on.

Methods proposed to support explainability are often divided into *ante-hoc* and *post-hoc* approaches. *Ante-hoc* approaches are related to systems that allow one to directly understand their mechanisms in providing a result such as a conclusion (e.g., a diagnosis) or a recommendation (e.g., a treatment option). Decision trees, rule-based models, and linear approximations are, for example, commonly considered to be implicitly explainable. Post-hoc approaches try to provide some explanation to the results reached by ML models, such as those based on deep neural networks, random forests, support vector machines, and many others. Post-hoc approaches are, in principle, applicable to different kinds of AI systems. The difference between these two approaches is that post-hoc approaches are not considered when designing a system, but deal with the extraction of explanatory information from an already existing system, which is usually based on ML "black-box" models. As we will see in this section, the distinction between post-hoc and ante-hoc approaches is sometimes subtle and has to be informed by further considerations.

Explainability is thus an inherently multifaceted concept, which still needs some more effort to have a precise characterization, also from the terminological point of view [1]. Let us now consider some dimensions of analysis that have been recently discussed in the literature.

The content of explanation: What is being explained? Independently from being either post-hoc or ante-hoc, XAI systems have to be specified and developed with respect to the subject of the provided explanation. Indeed, sometimes it is the reasoning mechanism itself that has to be explained. In this case, explanation focuses on the mechanics of the path that allowed the system to reach a specific result. Both generic and specific medical knowledge could be used to this regard. On the other side, explanatory information could be provided without any reference to the reasoning approach of the system, but focusing on deriving some form of association/relationship (causality) between data and corresponding results.

The stakeholders of explanation: Who needs explainability? Any kind of explanation needs to be tailored according to its recipients. It was recently highlighted that many possible stakeholders may be closely related to any XAI system [1]. In the medical and healthcare settings, among the possible stakeholders we consider a broad community of users, including clinicians, technicians, nurses, general practitioners, administrative staff, different kinds of students, healthcare policy makers, medical informaticians, and patients. The background knowledge of such stakeholders is often deeply different and often requires different user-centric solutions and techniques for a successful explanation.

The goal for explanation: Why is explainability required? Considering different stakeholders is not sufficient. We have to consider not only who is the recipient of the explanation, but also *why* the explanation is required. Indeed, the same stakeholder may have different motivations and requirements with respect to XAI systems. As an example, a physician may have different desiderata that include, variously, *education and experience*, *fairness*, *ethics*, *satisfaction*, *trust*, or *controllability*, while developers would consider *system acceptance*, possibly in addition to those required by a physician. Often such desiderata are not completely disjoint and may co-exist in a single XAI-system [1]. According to different desiderata, stakeholders could be looking for an answer to different questions related to explainability [2]: Why did the algorithm do that? Can I trust these results? How can I correct an error? Are data meaningful with respect to the required task?

The moment, the duration and the frequency of explanation: When, how long, and how frequently. A further, under-evaluated, issue is related to when and how frequently an explanation is requested of the system. Indeed, while naïve and occasional users often require frequent explanations at any stage of use of the supported AI system, experienced users who are supposed to use the system in the daily clinical routines, may require less frequent explanations, possibly focusing on rare or unexpected situations. The level of detail and thus the duration of the explanation may also be different, according to the specific needs of different stakeholders in different contexts, with different goals.

The modalities of explanation: How is explainability represented? Different choices are possible when deciding *how* to explain. A first option is to support *perceptive interpretability* [3]. This concept refers to interpretations that can be humanly perceived, (1) through the highlighting (often visual) of important input features with respect to a given output (*saliency*), (2) through the observation of the stimulation of neurons or groups of neurons (*signal interpretability*), and (3) through the composition of logical statements or sentences that can explain, even indicating causality (*verbal interpretability*). Often, perceptive interpretability is founded on an abstraction of the task at hand, which focuses on the most important aspects that explain the reached solution. Systems based on perceptive interpretability work with different techniques with respect to the ones used for the given task. For example, a fuzzy rule-based system may be coupled with an artificial neural network (ANN) system in diagnosing electrocardiographic (ECG) signals [4]. As for perceptive interpretability through visual and graphical systems, a widely acknowledged distinction exists between directly understandable data, which are visualized through one or two dimensional representations, and multi-dimensional representations, which are not directly understandable [2]. A second option is to consider *interpretability by mathematical structures*. In this case, either simple mathematical models are used, or different data-oriented approaches are used to highlight hidden features of data, such as data clustering, perturbations, data dependencies. Systems which support interpretability via mathematical structures consider outputs (which are ultimately perceptive) that require deeper cognitive processes and background knowledge, before being interpretable [3].

Further distinctions about the modalities of explanations supported by different XAI systems consider *model-agnostic approaches* and *model-specific approaches*. While the first approaches attempt to provide explanatory information only by observing input/output associations, model-specific approaches consider also specific features of the model under explanation [1]. A last aspect to consider for XAI systems is their *scope*. Indeed, some contributions focus on single predictions/classifications of the supported system (i.e., a single pair of inputs/output). Such systems have a *local scope* [5,6], in comparison with other approaches that have a *global scope*, which are designed to explain the overall reasoning mechanism of the model. Moving closer to applications in medicine, some aspects of AI have been identified that make XAI systems in medicine challenging but worthy of rigorous investigation. Factors as risk and responsibilities, accountability, and trustworthiness, even though already considered in non-medical domains, become here prominent and multifaceted. As an example, while explainability is a strong requirement in the clinical domain, as for acceptance, accountability, and legal compliance, a certain level of opaqueness can be acceptable for some clinical users, provided that some functional understanding of the model is supported, disregarding a possible low-level algorithmic understanding [2,3,7].

As XAI in medicine is in an early stage of investigation, some further issues have to be faced. Among them, the evaluation of XAI systems with actual end-users will help understand, represent, and satisfy user requirements [1]. *Causability* is the term proposed in [2] to explicitly highlight the need of measurements for the quality of explanations. In this direction, explanation interfaces have to make the results obtained through the explainable model both usable and useful to the considered stakeholder. Causability is thus a measure for the usability of such a human–AI interface.

All the previous arguments we discussed lead to a further, recently highlighted consideration [1]. Researching and developing XAI in medicine is an interdisciplinary task, which requires the active participation of different stakeholders, to cover different perspectives. Methodologies for the design of XAI systems in medicine would require skills from different scientific domains, such as AI, medical informatics, software engineering, medicine, healthcare, and cognitive sciences.

2.2. Applications

We find many examples in the literature of research activities that are devoted to exploring XAI in medical domains. Here we report some recent examples regarding different techniques.

ML algorithms such as neural networks are inherently non-explainable and are typically referred to as “black-box” models. However, there are some examples where neural network models can be shown to produce explanatory descriptions to support the interpretability of the output. In one study, the authors proposed a modular framework, CEFEs (CNN Explainability Framework for ECG signals), a post-hoc tri-modular evaluation structure that provides local interpretations and explanations from convolutional neural networks [8]. The evaluation of the model’s capacity is performed through quantitative interpretability, where the metrics represent the features learned by the model. In addition, the visualization of the features allows visually correlating the features. Pennisi et al. employed a novel lung-lobe segmentation network to identify CT scans of COVID-19 patients and automatically categorize specific lesions [9]. They integrate the pipeline into a web application to support radiologists in the investigation of this disease.

In recent years, ensemble learning has achieved excellent results incorporating explainability. Yeboah et al. present an ensemble clustering-based XAI model for traumatic brain injury (TBI) prognostic and diagnostic analysis [10]. The goal is to identify patient subgroups and key phenotypes that delineate these subgroups using tomography data, exploring the features’ relevance. In another example, the authors proposed an auxiliary decision support system that combined ensemble learning with case-based reasoning (CBR) to help physicians improve the accuracy of breast cancer recurrence prediction [11]. They use extreme gradient boosting (XGBoost) to predict the risk of breast cancer recurrence, and then use CBR to explain the reason for the prediction. Of note, they conducted a survey of 32 oncologists to assess the utility of the system as perceived by users, measuring the evaluation of the system through a questionnaire, leading to a positive assessment by the users of the system.

There are different examples of the usage of systems that exploit the explanations through rules-based systems extracted from medical data. They generated explanations in a human-understandable format, increasing the trust to believe the results given by the support system. El-Sappagh, et al., proposed a system of fuzzy IF-THEN rules [12]. It integrates reasoning with fuzzy reasoning over an ontology. They proposed and implemented a new semantically interpretable fuzzy rule-based system framework for diabetes diagnosis that is able to provide accurate decision support as a result. Kavya et al. developed an Allergy Diagnosis Support System (ADSS) [13]. They applied several ML algorithms and then selected the best-performing algorithm using k-fold cross-validation. In terms of the XAI method, they developed a rule-based approach by building a random forest. Each path in a tree is represented as an IF-THEN rule, and these rules are stored in a rule base for expert assessment. Additionally, the authors developed a mobile application, which can assist junior clinicians in confirming the diagnostic predictions.

Although the user represents a central aspect in the approaches we have just seen, the creation of an explainable system to use in a particular context requires a multi-disciplinary collaboration, involving collaboration with the stakeholders. Schoonderwoerd et al. presented a case study of an application of a human-centered design approach for AI-generated explanations [14]. The approach consisted of three components:

(i) Domain analysis to define the concept and context of explanations;
(ii) Requirements elicitation and assessment to derive the use cases and explanation requirements; and
(iii) The consequential multi-modal interaction design and evaluation to create a library of design patterns for explanations.

They apply this system in the context of child health. Dragoni, et al. proposed an XAI system based on logical reasoning that supports the monitoring of users’ behaviors and persuades them to follow healthy lifestyles [15]. In this case, the authors first assessed the usability of the application with questionnaires filled out by the user. Second, they validated the correctness of the explanation generated by the system. Finally, the last evaluation included an effectiveness analysis of the generated explanations.

3. Towards a foundational definition of XAI in medicine

We propose a conceptual framework for XAI that captures the intersection of four characteristics that are typical of any information system, statistical model, or software application. These characteristics are *Interpretability*, *Understandability*, *Usability*, and *Usefulness*, respectively. *Interpretability* is the degree to which a user can intuit the cause of a decision and thus the ability of a user to predict a system’s results [16]. *Understandability* is the degree to which a user can ascertain how the system works, and leads directly to user confidence in the system’s output. *Usability* is the ease with which a user can learn

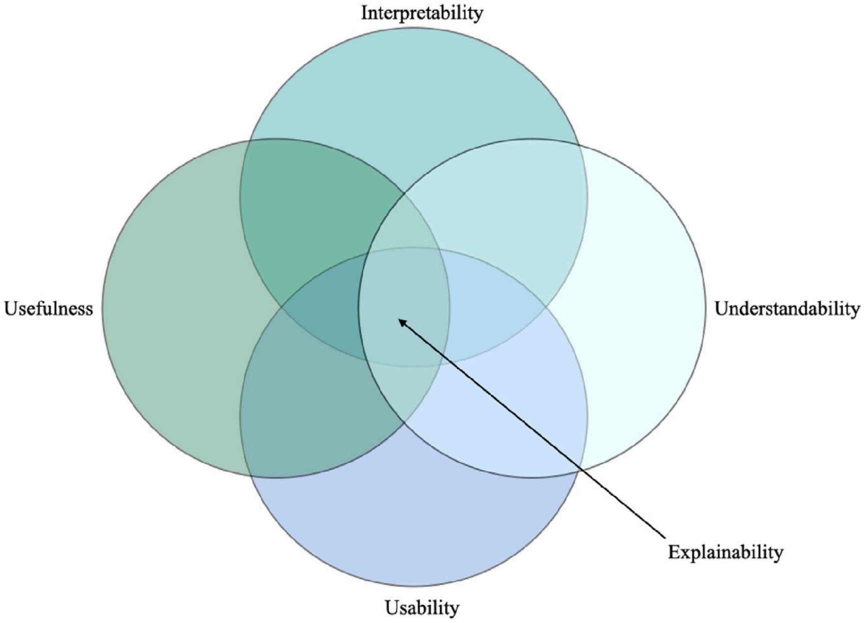

Fig. 1. The Venn diagram of explainability as intersection of usability, usefulness, interpretability, and understandability.

to operate, prepare inputs for, and interpret outputs of a system or component. Usability thus asks the question "Can one use the system easily?". *Usefulness*, on the other hand, asks the question "Will one use the system because it meets a user's needs?", and is seen as the practical worth or applicability of a system. A system is unlikely to be useful if it is not usable, however. As a result, usability is generally a first-order requirement of any information system or software application [17].

However, when it comes to AI, we are not talking about *any* information system. Rather, AI systems and applications typically realize some kind of reasoning task, to support some kind of decision-making, such as proposing a clinical diagnosis or controlling a task in an engineering operation, or to derive new knowledge/information in some specific context, as mining hidden patterns in patients' clinical histories. Perhaps unique to AI applications, we need two additional dimensions in order to realize an ability to provide user confidence that the decision was correct, but even more so, the ability for a user to ascertain how the system works. Thus, we propose that *understandability* is one such dimension, and furthermore that it is, in our framework, complementary to usability. That is, usability is enhanced via understandability: an AI application that is understandable is more likely to be usable.

The first characteristic of AI systems that we consider to be central to our framework is *interpretability*, which we construe as the degree to which a user can intuit the cause of a decision; in addition, it is the degree to which a human can consistently predict a model's results, based on her experience with the application. Just as understandability and usability are complementary, we propose that interpretability and usefulness are complementary as well. For example, a user of an AI application is more likely to find it useful, something that would meet her needs for a given purpose, if the result or decision made by the application is interpretable in the face of a real-life contingency.

This framework is illustrated as a Venn diagram in which these four characteristics overlap various points of articulation, but most importantly in the center, where all are needed when considering explainability, as is shown in Fig. 1.

Where these four characteristics intersect is that smallest, yet richest, segment of the Venn diagram, *explainability*. Due to the intersectionality of the four characteristics just described, explainability is a complex concept. It is not merely a characteristic of the model, but rather something that emerges from the intersection of the four characteristics we addressed here. As a result, we maintain that it is best to describe explainability in a multidimensional way through addressing a series of seven questions through the lens of others who have worked extensively in this domain.

The proposed foundational definition of XAI does not explicitly contain any specific reference to the medical and health domains. Indeed, the concepts introduced here are general and can be applied to any domain. However, we would stress here that, to the best of our knowledge, the definition of explainability as the intersection of four different characteristics is both original and particularly well suited for medicine and health AI.

As for the novelty of our definition, we identify here two different aspects: (i) from one general side, we explicitly distinguish the concepts of *interpretability*, *understandability*, and *explainability*. Such distinction is not clearly discussed in the existing literature, where, for example, *interpretable* and *explainable* are often taken as synonyms (see, for example, [1,3,18,19]). On the other hand, we explicitly introduced *usability* and *usefulness* as first principles of explainability. Such user-oriented aspects of explainability, even though considered and highlighted in the considered literature, have not been discussed as main component of a complex concept as complex as that of explainability.

The highlighted novelty of our foundational definition is also the leverage for making it especially well-suited for medicine and healthcare. Indeed, in our view, Medicine and healthcare are characterized by some specific features, which need to be considered as central for XAI. The first feature consists in the presence of *distributed, heterogeneous decision-making tasks* and a second can be defined as *knowledge-intensive domain*. The presence of *distributed, heterogeneous decision-making tasks* and of the corresponding XAI systems justify the presence of usability and usefulness in the definition of explainability. Indeed, usability and usefulness have to be evaluated according to different users and tasks. They are not absolute concepts and need to be assessed "on the field". The usability of systems that have to be adopted by specialized physicians in some intensive clinical setting requires it to be evaluated by the pertinent clinical stakeholders, while, for example, the usability of XAI systems supporting the communication and shared decision-making among clinicians, general practitioners, and patients (e.g., in a web app supporting the mental health monitoring of home patients) should be suitably assessed according to different explainability requirements, corresponding to different background knowledge and roles of the involved stakeholders.

Moreover, such *knowledge-intensive* and decision-intensive tasks require one to distinguish between interpretability and understandability. Indeed, while the concept of interpretability is related to the capability of predicting a system's result, even without being aware of the "internal" structure and functioning of the system, understandability refers to the capability of being aware of how the system works. In many intensive decision-based tasks, such as the prompt reaction to some unexpected change in an ICU patient's condition, the interpretability of an AI-based system may emerge as an indispensable feature. Indeed, the clinician has to be able to recognize how recorded vital signs are related to the alarms triggered by an AI-based system. It is worthwhile to stress that interpretability does not mean that the AI-based system is not important or useful as the user is able to predict the system's result. Indeed, the capability of predicting the system's result, does not mean that a human can process all the required data in an acceptable way, according to the requirements either related to the number of patients to consider or to the real-time results.

On the other hand, interpretability alone is often not sufficient to attain a necessary level of explainability. Understandability requires that the stakeholders have to be able to understand how the AI-based system works. In many medical and healthcare AI-based systems it may be important to have a deep understanding of the system internal behavior, in a way comprehensible to the specific clinical stakeholder. Let us continue with the example of an AI-based system for patient monitoring in ICU. While the AI-based system supporting real-time monitoring requires some kind of interpretability, the same AI-based system in the reporting and data analytics part could require more explicitly some kind of understandability. Indeed, when doing off-line data analysis it may be important to understand how the system is able to derive even unexpected results. As these results have to be related to existing and evolving medical knowledge, a deep comprehension of system technicalities and behaviors would also support a suitable elicitation of new medical knowledge.

4. Questions, propositions, and desiderata in the quest to attain XAI in medicine

After the proposal of our foundational definition of XAI in Medicine, supported by some simple examples in clinical domains, let us now move to more concrete issues that are necessary to consider in the practical development and use of (explainable) AI-based systems in medicine and healthcare. In the following we will touch on several different issues. After considering the design of XAI systems in Medicine (What are the requirements for XAI? How can we evaluate the goodness of the provided explanation?), we will introduce some further motivation supporting the distinction between understandability and explainability (If an AI system's output is understandable, is it automatically explainable?). Then, we will deal with the importance of modeling the considered medical domains (What is the role of domain understanding in achieving XAI in medical applications?). We will then continue with some more abstract aspects, as they relate to the evolution from data to wisdom through explainability (Can explainability draw us closer to wisdom?), to (Can an AI system that is not explainable be trustworthy?) and that connecting explainability and trustworthiness (Can an AI system that is not explainable be trustworthy?). We will end this section by answering the (usually hidden) question: Is XAI in medicine always required?

The questions we will deal with in this section complement the foundational definition we proposed in the previous section and apply such definition with respect to real-world aspects of XAI in clinical contexts.

4.1. What are the requirements for XAI? How can we evaluate the goodness of the provided explanation?

Proposition: There are tangible, instantiable, user-centered requirements that must be met in order to achieve an XAI system; more specifically, there is the need to measure, interpret, and understand usability vs. usefulness, and interpretability vs. understandability, and how those two relate to each other in the context of use and users, particularly in the context of AI in medicine. Similar to any information system, systems that employ AI can and should be developed and evaluated using state-of-the-art methods that can be extended to the domain of explainability. While validation and verification have been part of the canon for evaluating AI systems for several decades, these focus on operability and the accuracy of knowledge representation and inference. However, neither validation nor verification have fully taken into account the explainability or interpretability of the results from a user's perspective. Proposed here are desiderata in two broad domains of requirements for XAI that would serve to further the development of AI systems that help users to understand how such systems reach conclusions or offer advice. These domains are linking the cognitive to the explainable, and the evaluation of explainability.

- *Linking the cognitive to the explainable: the role of theory*. Knowledge elicitation has long been the central purpose of knowledge engineering, but it focuses on developing a knowledge base that does not address the needs of users as they interact with an AI system. This lacuna is especially evident with regard to the user interface. It is argued here, and supported in the literature, that *qualitative inquiry driven by theoretical frameworks* is needed to develop user-centered interfaces for ML in healthcare applications [20]. Theory-driven user interface design that takes into account the cognitive and behavioral aspects of users is foundational to achieving true explainability. This extends traditional principles of user interface design to include aspects of what influences user interpretation. Such aspects include attitudes and beliefs that may bias interpretability and subsequently influence users' confidence and understanding of the system and its results. In their recent survey of models for achieving explainability, Markus et al. provide a framework for choosing the type of explainable interface between model-, attribution-, or example-based explanations [19]. They advocate for methods for achieving explainability that are sensitive to the requirements of the problem domain, and that these should drive the choice of approach, rather than enforcing a single paradigm of explainability. In a word, they call for an *agile* approach to attaining and evaluating explainability, which is very much in line with the best practices of information system development in general. One agile approach to attaining explainability in AI systems turns to fuzzy set theory and its application to fuzzy reasoning systems. Such systems provide a plausible paradigm for modeling explainability, since natural language is one defining characteristic of fuzzy systems. Alonso Moral et al. argue for this paradigm, showing how user-centered explainability is connected to fuzzy modeling [21]. Finally, any effort to establish explainability needs to be linked to the cognitive aspects of human inference. It is arguable that there is no more urgent need for this in medical decision making. An example of this kind of cognition is seen in the principle of *ex adiuvantibus*, which is the inference leading to a conclusion, such as the cause of a diseases, that is based on evidence that the disease responded to a treatment. As an example, one might infer that a migraine headache was caused by exposure to a specific allergen because an antihistamine was shown to prevent the headache. Such causal inferences many or may not be correct in practice, but they are made frequently in clinical practices, and in fact this type of reasoning is at the heart of allopathic medicine.

- *A user-oriented perspective of explainability*. The growing research community in XAI has already developed a number of highly successful XAI methods [22]. Explainability in this context highlights technically decision-relevant parts of machine representations and machine models. For example, parts that contributed to model accuracy during training or to a particular prediction are visualized by a heatmap, a good and proven example being the very well known Layer Wise Relevance Propagation (LRP) method [23]. However, this visualization does not refer to a human model. For this purpose, the concept of causability was introduced, which is defined as the measurable extent to which an explanation reaches a certain level of causal understanding for a human end-user [2]. Since this concept refers to a human model, it can be used very well to design and evaluate future human–AI interfaces [7]. These future Human–AI interfaces must provide a successful mapping between Explainability and causability and foster contextual understanding and allow the expert to ask questions and counterfactuals ("what-if" questions) [24]. At the same time such question–answer interfaces can make use of a human-in-the-loop, who can bring human experience and conceptual knowledge to AI processes — something that the best AI algorithms available still lack. An example that is important for medical AI is the classification of entities into several classes, where typically, taking into account the uncertainty about the membership of the classes, entities are classified as "yes", "no", or "maybe". However, in doing so, it is desirable – especially in medical problems – to indicate the propensity or probability of a classification to belong to a single yes or no category. Neural networks have proven their high performance in crisp classification, however, as we know, the solution is not comprehensible and therefore difficult or impossible for a human expert (e.g. a physician) to interpret and understand. Rule-based systems are in principle explainable, however they are based on formal inference structures and also have problems with interpretability due to their high complexity. We must emphasize that even human experts sometimes cannot explain, but construct mental models of the problem and use these models to select the best possible solution. Hudec et al. propose a classification by aggregation functions of mixed behavior through the variability of ordinal sums of conjunctive and disjunctive functions [25]. In this way, domain experts should assign only the most relevant observations regarding the considered attributes. Consequently, the variability of the functions provides room for ML to learn the best possible option from the data. Such a solution is tractable, reproducible and explainable to domain experts.
- *Evaluating explainability*. Ultimately, explainability is in the eye of the beholder, i.e., the user. As such it is incumbent on those who aim to develop XAI systems to account for their usability, but also their usefulness. *Usability* can be measured using modifications of such instruments as the System Usability Scale (SUS). A detailed retrospective examination of the SUS is provided in [26]. Modification to this scale would need to account for the interpretability of the system, including both inputs and outputs. Another approach to usability assessment is one that focuses on causality [2,27]. This approach allows users and developers to trace inferential pathways and evaluate them for plausibility. As such, not only can inferential errors be identified rapidly, the reasoning behind them can, as well. Using this scale, a deep assessment of usability can be obtained throughout the system development life cycle. Yet another approach to assessing usability focuses on user-centered reporting of results, such that users provide important input on and influence over what is reported by the system. This was shown to be an effective way to ensure that random forest results were reported in a way that users found them to be interpretable [28]. However, none of these approaches to evaluating explainability address the issue of *usefulness*. While a system may be usable, it is not necessarily useful, meaning that the system addresses some important task, telling a user something they did not already know or infer from available facts or knowledge. To assess usefulness, one needs to turn to long-term, post-hoc qualitative and quantitative evaluation of how, when, and why the system is being used and in what contexts does it fit (or fail to fit) workflows. Another consideration for usefulness is whether or not a system is used in practice to replace another. This is especially important in busy clinical settings, where AI systems might be used to augment medical decision making. However, if a system is not useful, practitioners will not use it, even though it might be very usable, or they will use the system but develop workarounds to make it more useful, sometimes with consequences that are potentially catastrophic to patients. For this type of evaluation, the frameworks mentioned above can inform the development of strategies and methods for observing the use of AI systems in these contexts in real-time, and the framework-driven analysis of data obtained during this endeavor.

4.2. If an AI system's output is understandable, is it automatically explainable?

Proposition: Understanding the output from an AI system is foundational to explainability, but it is only one requirement that has to be merged with usability, usefulness, and interpretability to compose explainability. A central goal of ML is to build a model which summarizes linear and/or nonlinear patterns in a dataset. Good models are useful for making predictions in new data and thus have the quality of generalizability, which in turn makes them useful. Most ML models, such as those derived from neural networks or gradient boosting, have an underlying mathematical foundation. For example, a neural network model can be written as a summation of products of weights and inputs from data and hidden layer nodes. Thus, our knowledge of the mathematical foundation of a model makes it inherently understandable in that we know the function that relates the data inputs to the outcome being predicted. Our understanding can be improved by conducting experiments on the model by, for example, perturbing inputs and/or model components to observe their effects on model quality metrics. We can even decompose the model into linear and nonlinear components using these kinds of perturbation experiments when combined with entropy-based measures from information theory, for example. In this way, it is possible to gain a good understanding of a model. But does understanding translate to explainability?

As previously described, characteristics of XAI include usability, usefulness, interpretability, and understandability. Knowing the mathematical basis of a model does not necessarily make it useful. For example, a neural network model might do a good job of predicting 30-day hospital readmissions following surgery. Further, the model might generalize well to clinical data from other hospitals. The model is understandable because the mathematical basis is known and can be described. Although the model is predictive and understandable, it might not be useful for reducing readmissions if the features include patient demographics such as gender and zip code which cannot be changed to improve the outcome. As another example, consider a neural network model relating gene expression features to risk of disease, where the predictive features include a number of housekeeping genes required for the maintenance and function of all cells. The model might be understandable and useful, but it might not be interpretable. In other words, it may be difficult for the domain expert to come up with an explanation for why this set of genes contributes to disease risk when they impact every cell in the body. This in turn would limit the ability of a pharmacologist to develop a therapeutic intervention.

Understanding an ML model is thus a first step towards XAI. While complementary, usability, usefulness, interpretability, and understandability can be synergistic. For example, a domain-specific knowledge graph can make a model more understandable and more interpretable by informing the user of biological relationships among the features [29]. Further, biomedical ontologies can facilitate both understanding and interpretation because the feature relationships have been described through a synthesis of multiple knowledge sources that capture their semantic meaning [30].

4.3. What is the role of domain understanding in achieving XAI in medical applications?

Proposition: XAI-based systems need to start from modeling the biomedical and clinical domain in order to obtain a true understanding of the context in which these systems will be used. As stated by several authors, a key aspect of building biomedical (and in particular clinical) AI-based systems is to understand the context. For example, understanding the context of clinical decisions means to model the patients' careflow: identify the key actors of care and the decision-makers, explicitly define the timing of decisions, and clarify the data collection phases and their critical elements, including the potential sources of missing data. Only by deeply analyzing all these aspects it will be possible to design a successful AI-based system and to properly identify the explainability components. The real importance of an AI system in medicine is to support the planning and delivery of medical treatment more than just perform diagnostic labeling [31]. XAI is essential to achieving this goal, in addition to the strategies to induce trust in AI-supported decisions.

To this end, there is the need for integrating stakeholders and users into entire AI development life-cycle. Following the approach proposed by Bellazzi and Zupan in [32], a potential strategy is to apply in the design of AI-based systems the same conceptual model proposed for data mining models by the Cross Industry Standard Process for Data Mining (*CRISP-DM*) process model. CRISP-DM has six phases that are helpful to obtain explainable systems "by design":

1. Business understanding;
2. Data understanding;
3. Data preparation;
4. Modeling;
5. Evaluation;
6. Deployment.

While data preparation, modeling, and evaluation are now reported in all ML textbooks, very often little attention is given to business understanding, data understanding, and finally deployment. All of those are related to understanding the biomedical context, modeling the process and clearly expressing the goals. Data needs to be modeled; as well, it should be understood who and when data are collected, which is often related to the nature of missing data. Finally, having clearly in mind the deployment scenario is a key driver for designing XAI approaches. In this phase all actors involved in decision making should be involved, resorting to different instruments, from formal questionnaires to qualitative interviews. Several other development methodologies could be suitably adopted/extended/adapted when designing and implementing XAI systems in medicine, where the different stakeholders and the application domain are explicitly dealt with. As an example, well established methodologies as CommonKADS, supporting the design of knowledge-intensive systems coupled with UML notations, as well as methodologies dealing with the design of ontology- and/or data-based reasoning/analytics systems could provide suitable techniques for domain understanding and modeling [33–36].

As also reported by the EU white paper [37], AI systems and their decisions should be explained in a manner that is adapted to the appropriate stakeholder.

Among the specific features of medicine and health, we have to consider when designing XAI medical applications, we distinguish here:

- *The heterogeneous nature of medical data.* Medical data consists of images, movies, biosignals, and structured and unstructured alphanumeric data from electronic medical records. All of this data needs to be suitably integrated into and consistently and appropriately managed by XAI medical applications. Even though explainability has been considered for these different kinds of information systems (see, for example [38,39]), further research efforts will have to deal with the elicitation of both visual and textual knowledge from such kinds of data, often left partially implicit by skilled physicians [40]. As an example, while radiology is mainly based on images, which are visually analyzed by radiologists even by the support of computerized devices, and related natural language reports, oncology deals mainly with knowledge represented in a textual way, often highly structured (as in the case of chemotherapy guidelines), while cardiology has a lot of information and related knowledge expressed through biosignals (e.g., the electrocardiogram) and movies (e.g., echocardiograms).
- *The presence of highly specialized knowledge in different clinical and healthcare domains*. Specific domains as cardiology, oncology, neurology, healthcare policy, and so on, have their own vocabulary, specific shared knowledge about diagnosis, treatments, and so on [41,42]. XAI systems have, thus, to deal with jargon, abbreviations and terminological heterogeneity, idiosyncratic usage habits, and different kinds of knowledge, as previously stressed, especially when they have to support the exchange of shared information [43].
- *The presence of many different specialized processes, requiring the coordination of different stakeholders*. Explainability in medicine and health is often related to the results of prediction and/or classification tasks towards diagnosis and/or therapy effects and so on. Besides this "static" part, clinical tasks as monitoring, diagnosis, therapy, and prognosis are merged in a "dynamic" context, composed of complex medical or healthcare processes and pathways. In such processes, different healthcare actors, as clinicians, epidemiologists, nurses, and technicians, are involved with different roles. XAI systems cannot avoid facing these intertwined aspects, related to knowledge, information, processes, and actors, to suitably support specific clinical activities [44].

4.4. Can explainability draw us closer to wisdom?

Proposition: Explainability is a requirement to completing the data-information-knowledge-wisdom spectrum. Understandability is an essential prerequisite for the transition from information to knowledge and provides a path to the realization of knowledge as wisdom. Explainability can, on the one hand, promote trust on the part of end users (compare with the previous section), and, on the other hand, promote understanding and, in turn, trust on the part of developers of algorithms, and finally also provide new insights. Trust is of eminent importance and is often underestimated and in order to bring AI into the real world, it must be trustworthy [45]. To be trustworthy, any AI must comply with applicable rules and regulations, adhere to ethical principles [46], follow legal issues [47] and be implemented in a secure and robust manner. This is particularly required by the EU High-Level Expert Group on AI.[1]

In classical philosophy since ancient Greece, explanations have always been central, as the word philosophy itself means "love of wisdom". A good example is the deductive-nomological model of Hempel and Oppenheim (1948) [48] which is based on a formal structure of scientific explanation of a causal relationship using natural language. The model consists of two parts, the proposition to be explained (explanandum) and the explanation itself (explanans), which is composed

[1] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (access: February, 09, 2022).

of general law statements and (empirical) boundary conditions (antecedent statements) as premises. The preliminary work on this was already served by Karl Popper in his work “Logic of Research” [49].

Colloquially, explanations differ in their completeness or degree of causality [50]. In his work, Tim Miller (2019) [51] combined insights from the social sciences with explanations in AI and divided explanatory questions into three classes: (1) what-questions, such as “What event happened?”; (2) how-questions, such as “How did this event happen?”; and (3) why-questions, such as “Why did this event happen?”.

“Moving closer to wisdom” implies also that physicians and other clinical stakeholders receive some feedback on their own capabilities and attitudes towards explainability: do we need that AI systems have sophisticated explainability capabilities when it happens that physicians do not spend any effort to explain their choices? From one point of view, we could say that requirements about explainability have to be more strict for AI systems. Indeed, “we may hold physicians responsible for their lack of explainability and potential mistakes,but we cannot do the same with AI” (from [52]). On the other side, XAI systems can support clinicians in providing even more sound and founded decisions. In this direction, AI systems have to be considered as tools that require a specific and sound certification process also with regards to explainability. Similarly to what happens to marketed drugs, which need to follow strict certification processes before being approved, also XAI systems should be formally approved before used in real world clinical and healthcare contexts. Such kind of approach, followed by a continuous monitoring after the introduction of such tools in real clinical contexts, would help to clarify responsibilities both for physicians and for the producers of XAI systems.

4.5. Can an AI system that is not explainable be trustworthy?

Proposition: XAI is an integral component of trustworthy AI systems. In 2019 the EU has published the Ethics Guidelines for trustworthy AI, which contains a general framework where explainability represents an important component.[2] These guidelines have been used as a basis for some of the sections of the proposal of the Artificial Intelligence Act released by the European Commission in April 2021. The guidelines correctly states that “Trust in the development, deployment and use of AI systems concerns not only the technology’s inherent properties, but also the qualities of the socio-technical systems involving AI applications ... it is not simply components of the AI system but the system in its overall context that may or may not engender trust”. To this end, AI systems should be lawful, i.e., complying with laws and regulations, ethical, i.e., being to ethical principles and robust, both from a technical and social perspective. The guidelines also provides seven requirements for implementation of AI trustworthy solutions, including:

- human agency and oversight
- technical robustness and safety
- privacy and data governance
- transparency
- diversity non-discrimination and fairness
- societal and environmental well-being
- accountability.

Explainability is considered as a component of transparency, together with traceability and communication. In our view explainability has an horizontal impact which is wider than what is stated in the guidelines. First of all, within transparency, it has a strong overlap with communication, which is related to understandability. Second, explainability is a key component of accountability, since it provides instrument to keep track of the decisions, going back to the “reasons-why” an AI tool, or a decision-maker empowered by AI solutions, has suggested the decision. Finally, it can be considered as a way to ensure technical robustness, providing explanations about change in decisions related to changes in the attribute values; this provides ways to control the performance of the algorithms and identify aberrant situations. Rather interestingly, trustworthiness allows to jointly consider two related concepts: explainability and reliability. “Reliability” is a component of robustness that indicates the degree of trust that we have on the prediction made by an ML model on a single example [53]. Coupled with local explainability ensures that local predictions can be used in a safety critical context as medicine is.

4.6. Is XAI in medicine always required?

Proposition: Explanations are not always required in order for an AI model to be useful. Functional specifications obtained from deep analysis of the problem domain and users should determine when explainability and interpretability are required. While many recognize the necessity to incorporate explainability features in AI models, addressing user needs for understanding AI remains an open question. As the type of interpretability needed varies depending on the context, it is clear that XAI must take a human-centered approach. The same explanation may be more or less comprehensible to different users or even to the same user engaged in different roles and we should not confuse the different notions of interpretability because each kind serves a different purpose [54]. For instance, we cannot provide algorithm designers and end users with the same explanations. An ML expert might prefer an explanation that helps them debug the model and understand its inner-working [55]. In contrast, an end user might require a causal explanation of predictions to ensure that decisions informed by those predictions are fair [56].

The use of techniques to explain AI models has become central in human-centered systems. For example, visual analytics systems help users understand and interact with AI models by providing them with visualizations and tools that facilitate the exploration, analysis, interaction with AI models. To close the gap between XAI methods and user needs for transparency, the human–computer interaction community has called for interdisciplinary collaboration [57] and user-centered approaches to explainability [58]. The need to create effective explainability features in diverse medical applications led to novel ways to probe user needs. As an explanation can be seen as an answer to a question, Liao et al. represented user needs for explainability in terms of questions a user might ask about the AI model, thus creating a question bank, a list of prototypical user questions that XAI methods can address [59]. It is essential that model developers understand why an explanation is needed and what type of explanation is helpful for a given situation.

AI models do not need to be interpretable to be useful [60]. In this context, a blanket rejection of black-box methods in decision support systems may be hasty. For example, suppose an AI model yields accurate predictions that help clinicians better treat their patients. In that case, it may be useful even without a detailed explanation of how or why it works. Therefore, it is essential to identify biomedical applications in which black-box answers generated by AI models can have a useful role in decision support systems and thus can be safely used.

When an AI model produces the best results or yields accurate predictions that help clinicians better treat patients, it may be useful even without detailed explanations. For example, in reading medical images, trained AI systems enhance the performance of human radiologists in detecting cancers [61,62]. That is not to say that AI interpretability is not valuable. In particular, when AI models are used in an automated fashion, laws and regulations should require a causal explanation of AI decisions to ensure that they are fair [63]. However, in situations when AI models do not lead to automated decision making, an explanation

[2] https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

may not be needed and auditing [64] together with judicious testing of AI models via randomized control trials [65] might be sufficient.

Although the process used by AI models to generate predictions can be limited and biased, it is also different from human thought processes in ways that can reveal new connections. This creates a case for using black-box AI models as tools to guide human inquiry [66,67]. For example, in a groundbreaking medical imaging study, a deep learning model was trained to diagnose diabetic retinopathy from retinal images [68]. The model achieved performance comparable to a committee of ophthalmologists. Further, the model accurately identified several characteristics that are not generally assessed with retinal images, including cardiological risk factors, age, and gender [69]. No one had previously noticed gender-based differences in human retinas, so the black-box observation inspired researchers to investigate how and why male and female retinas differ.

Moving to a final example, XAI needs to be declined in different ways in different contexts. Indeed, the explanation requirements regarding clinical medicine, for example, may have to deal with specialized physicians, who could have a knowledge in the specific domain that not requires XAI (but an AI system with certified good performances), while, considering, for example, the issue of pandemic management, requirements from epidemiology or national health policies and management could be extremely demanding, as possible relevant public decisions have to be suitably justified [70].

5. Conclusions and research directions

The issue of explainability in AI is evolving at a rapid pace. As we have seen in this paper, there has been considerable research into XAI, but there is still much to be done. We note here five broad areas where more research is needed.

- **Bridging the gap between symbolic (ante hoc) and sub-symbolic (black-box) approaches.** Sub-symbolic ML approaches and symbolic ones are currently considered by two research communities, having often completely different perspectives and background. XAI requires that such dichotomy has to be overcame. Indeed, symbolic approaches, as the ones related to logics-based proposals, ontologies, query systems, Bayesian networks, and so on, would be grounded in order to use them in establishing explainability [71]. Research on the seamless proposal of "hybrid" systems, merging both sub-symbolic and symbolic approaches still requires a lot of joint efforts.
- **Engineering explainability into intelligent systems.** An important, even fundamental question is whether and how explainability can actually be engineered into AI. Even given our conceptual framework for thinking about XAI (Section 2), we still need to address the idiosyncrasies of individual intelligent systems as well as those of their users. We contend that more specialized research into the structural, functional, and behavioral characteristics of these systems and the environments in which they are situated should be the targets of rigorous mixed-methods research that encompasses the entire system ecology, from *in silico* to *in vivo* contexts.
- **Evaluating and improving the effects of explainable components and approaches.** The evaluation of intelligent systems, as a scientific and methodological discipline, is changing, yet there needs to be more systematic investigation and implementation of these methods. Too often, there is emphasis on the accuracy of a decision made by such systems, typical as a proportion — whether in terms of overall accuracy (percent "correct"), or more nuanced indicators such as sensitivity, specificity, predictive values, or their derivatives such as the F-score or areas under the receiver operating characteristic curve or under the precision–recall curve. None of these well-used metrics indicate anything about the effects of the system on user beliefs, attitudes, or behavior. These are effects that require, again, deep mixed-methods research, this time applied to evaluating the effects of XAI (or its absence) on such issues as user acceptability, actions taken (or not) based on the results offered by the system, and overall impact on clinical or other workflows.
- **Determining when explainability is needed.** Is explainability always needed? This is a fair question, indeed. AI systems (or "subsystems" that work in the background to provide some inference to assist another systems might not require real-time explainability. A feature selection algorithm as part of a knowledge discovery or decision making workflow is one example if such an AI. However, we would argue that in order for software developers who need to use such subsystems in their work, explainability to them is critically important. They have to know how that subsystem works and why. But to the end-user of that workflow, it might not be so important in real-time. Rather, an in-depth description of the entire workflow and its components should be provided so the end-user understands how the overall system works. This situation begs the question again: "Is explainability always needed?". The answer, we propose here, is yes, but titrated to the needs of the user at particular times or in response to specific events.
- **Investigating the design of user-centered and user-tailored explainability artifacts.** If there were ever a more urgent need for rigorous research into user-centered design, it is hard to see one that surpasses the field of XAI. Such design, as noted, must be sensitive to workflow contexts, certainly, but there are other equally important considerations. One of the most important of these is the involvement of users into the design process. Rapid prototype design paradigms should be used in order to keep users involved during all phases of the development and implementation of AI systems. We already do this to some extent in the field of knowledge engineering, although history is full of examples where gaps in knowledge acquisition and representation have led to system failures, some with catastrophic results

We hope that our examination of the issues involved in developing XAI systems – our *manifesto*, if you will – will not be construed as the definitive work in this area. Rather, hope that the issues we considered here will stimulate further thought and hopefully fruitful research and development of XAI systems, particularly in medical contexts, but extending beyond to other contexts as well.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

To avoid any bias in the review process, this paper has been managed by the guest editors of the AIIM special issue "Explainable artificial intelligence and real-world applications in healthcare". At the end of the review process, we would like to thank Gregor Stiglic, Jose Juarez, and Peter Haddawy, who provided important comments and suggestions, allowing us to greatly improve the manuscript.

A.H. is partially funded by the Austrian Science Fund (FWF), Project: P-32554 "A reference model of explainable Artificial Intelligence for the Medical Domain". M.Z. is supported, in part, by National Science Foundation (NSF), USA under nos. IIS-2030459 and IIS-2033384, US Air Force Contract No. FA8702-15-D-0001, Harvard Data Science Initiative, USA, Amazon Research Award, USA, Bayer Early Excellence in Science Award, USA, AstraZeneca Research, United Kingdom, and Roche Alliance with Distinguished Scientists Award, USA. R.B. is partially funded by the project Periscope, USA (Pan-European Response to the ImpactS of COVID-19 and future Pandemics and Epidemics), ID:101016233, funded by the European Union. J.H. is partially funded by National Institutes of Health (USA) Clinical Translational Science Award, UL1-TR001878. J.M. is partially funded by National Institutes of Health (USA) grant AG066833. B.A. and C.C. are partially funded by the Ministry of University and Research, MIUR, Project Italian Outstanding Departments, 2018–2022. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

References

[1] Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, et al. What do we want from explainable artificial intelligence (XAI)? - A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 2021;296:103473. http://dx.doi.org/10.1016/j.artint.2021.103473.

[2] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and explainability of artificial intelligence in medicine. WIREs Data Min Knowl Discov 2019;9(4). http://dx.doi.org/10.1002/widm.1312.

[3] Tjoa E, Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. IEEE Trans Neural Netw Learn Syst 2021;32(11):4793–813. http://dx.doi.org/10.1109/TNNLS.2020.3027314.

[4] Bozzola P, Bortolan G, Combi C, Pinciroli F, Brohet C. A hybrid neuro-fuzzy system for ECG classification of myocardial infarction. In: IEEE conference on computers in cardiology. 1996, p. 241–4, cited By 23. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030420077&partnerID=40&md5=1e58bbe69e4c64e1291e4beea104ba58.

[5] Adhikari A, Tax DM, Satta R, Faeth M. LEAFAGE: Example-based and feature importance-based explanations for black-box ML models. In: 2019 IEEE international conference on fuzzy systems. IEEE; 2019, p. 1–7.

[6] Ahn S, Kim J, Park SY, Cho S. Explaining deep learning-based traffic classification using a genetic algorithm. IEEE Access 2020;9:4738–51.

[7] Holzinger A, Mueller H. Toward human-AI interfaces to support explainability and causability in medical AI. IEEE Comput 2021;54(10):78–86. http://dx.doi.org/10.1109/MC.2021.3092610.

[8] Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B. CEFEs: A CNN explainable framework for ECG signals. Artif Intell Med 2021;115:102059.

[9] Pennisi M, Kavasidis I, Spampinato C, Schinina V, Palazzo S, Salanitri FP, et al. An explainable AI system for automated COVID-19 assessment and lesion categorization from CT-scans. Artif Intell Med 2021;118:102114.

[10] Yeboah D, Steinmeister L, Hier DB, Hadi B, Wunsch DC, Olbricht GR, et al. An explainable and statistically validated ensemble clustering model applied to the identification of traumatic brain injury subgroups. IEEE Access 2020;8:180690–705.

[11] Gu D, Su K, Zhao H. A case-based ensemble learning system for explainable breast cancer recurrence prediction. Artif Intell Med 2020;107:101858.

[12] El-Sappagh S, Alonso JM, Ali F, Ali A, Jang J-H, Kwak K-S. An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access 2018;6:37371–94.

[13] Kavya R, Christopher J, Panda S, Lazarus YB. Machine learning and XAI approaches for allergy diagnosis. Biomed Signal Process Control 2021;69:102681.

[14] Schoonderwoerd TA, Jorritsma W, Neerincx MA, 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.

[15] Dragoni M, Donadello I, Eccher C. Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice. Artif Intell Med 2020;105:101840.

[16] Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, Tengg-Kobligk Hv, et al. On the interpretability of artificial intelligence in radiology: Challenges and opportunities. Radiol Artif Intell 2020;2(3):e190043. http://dx.doi.org/10.1148/ryai.2020190043, PMID: 32510054.

[17] Landauer TK. The trouble with computers: usefulness, usability, and productivity. MIT Press; 1995.

[18] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2019;51(5):93:1–42. http://dx.doi.org/10.1145/3236009.

[19] Markus AF, Kors JA, Rijnbeek PR. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inform 2020;103655.

[20] Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. BMC Med Inform Decis Mak 2020;20(1):1–16.

[21] Mencar C, Alonso JM. Paving the way to explainable artificial intelligence with fuzzy modeling. In: International workshop on fuzzy logic and applications. Springer; 2018, p. 215–27.

[22] Zhou J, Gandomi AH, Chen F, Holzinger A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics 2021;10(5):593. http://dx.doi.org/10.3390/electronics10050593.

[23] Montavon G, Samek W, Müller K-R. Methods for interpreting and understanding deep neural networks. Digit Signal Process 2018;73(2):1–15. http://dx.doi.org/10.1016/j.dsp.2017.10.011.

[24] Holzinger A, Malle B, Saranti A, Pfeifer B. Towards multi-modal causability with graph neural networks enabling information fusion for explainable AI. Inf Fusion 2021;71(7):28–37. http://dx.doi.org/10.1016/j.inffus.2021.01.008.

[25] Hudec M, Minarikova E, Mesiar R, Saranti A, Holzinger A. Classification by ordinal sums of conjunctive and disjunctive functions for explainable AI and interpretable machine learning solutions. Knowl Based Syst 2021;220:106916. http://dx.doi.org/10.1016/j.knosys.2021.106916.

[26] Brooke J. SUS: A retrospective. J Usability Stud 2003;8(2):29–40.

[27] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS). In: KI-künstliche intelligenz. Springer; 2020, p. 1–6.

[28] Petkovic D, Altman R, Wong M, Vigil A. Improving the explainability of random forest classifier–user centered approach. In: Pacific symposium on biocomputing 2018: proceedings of the pacific symposium. World Scientific; 2018, p. 204–15.

[29] Mensio M, Bastianelli E, Tiddi I, Rizzo G. Mitigating bias in deep nets with knowledge bases: The case of natural language understanding for robots. In: AAAI spring symposium: combining machine learning with knowledge engineering (1). 2020, p. 1–9.

[30] Confalonieri R, Weyde T, Besold TR, Martín FMdP. Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks. 2019, arXiv:1906.08362.

[31] Adler-Milstein J, Chen J, Dhaliwal G. Next-generation artificial intelligence for diagnosis: From predicting diagnostic labels to "wayfinding". JAMA 2021. http://dx.doi.org/10.1001/jama.2021.22396.

[32] Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. Int J Med Inform 2008;77(2):81–97. http://dx.doi.org/10.1016/j.ijmedinf.2006.11.006.

[33] Brachman RJ, Levesque HJ. Knowledge representation and reasoning. Elsevier; 2004, URL http://www.elsevier.com/wps/find/bookdescription.cws_home/702602/description.

[34] Nemati HR, Steiger DM, Iyer LS, Herschel RT. Knowledge warehouse: an architectural integration of knowledge management, decision support, artificial intelligence and data warehousing. Decis Support Syst 2002;33(2):143–61. http://dx.doi.org/10.1016/S0167-9236(01)00141-5.

[35] Schreiber G, Akkermans H, Anjewierden A, de Hoog R, Shadbolt N, Van de Velde W, et al. Knowledge engineering and management: the commonKADS methodology. Cambridge, MA: MIT Press; 2000.

[36] Vaisman AA, Zimányi E. Data warehouse systems - design and implementation. Data-centric systems and applications, 2nd ed.. Springer; 2022, http://dx.doi.org/10.1007/978-3-662-65167-4.

[37] European Commission. White paper on artificial intelligence: a European approach to excellence and trust. COM(2020) 65 final, Brussels: European Commission; 2020, https://ec.europa.eu/info/files/white-paper-artificial/-intelligence-european-approach-excellence-and-trust_en.

[38] Jin W, Li X, Hamarneh G. Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? In: Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22 - March 1, 2022. AAAI Press; 2022, p. 11945–53, URL https://ojs.aaai.org/index.php/AAAI/article/view/21452.

[39] Payrovnaziri SN, Chen Z, Rengifo-Moreno P, Miller T, Bian J, Chen JH, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. J Am Med Inform Assoc 2020;27(7):1173–85. http://dx.doi.org/10.1093/jamia/ocaa053.

[40] Holzinger A. Explainable AI and multi-modal causability in medicine. I-Com 2021;19(3):171–9. http://dx.doi.org/10.1515/icom-2020-0024.

[41] Powsner SM, Costa J, Homer RJ. Clinicians are from mars and pathologists are from venus: Clinician interpretation of pathology reports. Arch Pathol Lab Med 2000;124(7):1040–6. http://dx.doi.org/10.5858/2000-124-1040-CAFMAP, arXiv:https://meridian.allenpress.com/aplm/article-pdf/124/7/1040/2724930/0003-9985(2000)124_1040_cafmap_2_0_co_2.pdf.

[42] Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, et al. A natural language processing system that links medical terms in electronic health record notes to lay definitions: System development using physician reviews. J Med Internet Res 2018;20(1):e26. http://dx.doi.org/10.2196/jmir.8669.

[43] Rau NM, Basir MA, Flynn KE. Parental understanding of crucial medical jargon used in prenatal prematurity counseling. BMC Med Inform Decis Mak 2020;20(1):169. http://dx.doi.org/10.1186/s12911-020-01188-w.

[44] Combi C, Oliboni B, Zardini A, Zerbato F. A methodological framework for the integrated design of decision-intensive care pathways - an application to the management of COPD patients. J Heal Inform Res 2017;1(2):157–217. http://dx.doi.org/10.1007/s41666-017-0007-4.

[45] Holzinger A, Dehmer M, Emmert-Streib F, Cucchiara R, Augenstein I, Del Ser J, et al. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Inf Fusion 2021;79(3):263–78. http://dx.doi.org/10.1016/j.inffus.2021.10.007.

[46] Mueller H, Mayrhofer MT, Veen E-BV, Holzinger A. The ten commandments of ethical medical AI. IEEE Comput 2021;54(7):119–23. http://dx.doi.org/10.1109/MC.2021.3074263.

[47] Stoeger K, Schneeberger D, Holzinger A. Medical artificial intelligence: The European legal perspective. Commun ACM 2021;64(11):34–6. http://dx.doi.org/10.1145/3458652.

[48] Hempel CG, Oppenheim P. Studies in the logic of explanation. Philos Sci 1948;15(2):135–75.

[49] Popper K. Die logik der forschung. Zur erkenntnistheorie der modernen naturwissenschaft. Wien: Springer-Verlag; 1935.

[50] Pearl J. The seven tools of causal inference, with reflections on machine learning. Commun ACM 2019;62(3):54–60.

[51] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence 2019;267:1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.

[52] Kempt H, Heilinger J, Nagel SK. Relative explainability and double standards in medical decision-making. Ethics Inf Technol 2022;24(2):20. http://dx.doi.org/10.1007/s10676-022-09646-x.

[53] Nicora G, Rios M, Abu-Hanna A, Bellazzi R. Evaluating pointwise reliability of machine learning prediction. J Biomed Inform 2022. http://dx.doi.org/10.1016/j.jbi.2022.103996.

[54] Weller A. Transparency: Motivations and challenges. In: Explainable AI: interpreting, explaining and visualizing deep learning. Springer; 2019, p. 23–40.

[55] Ying R, Bourgeois D, You J, Zitnik M, Leskovec J. GNNexplainer: Generating explanations for graph neural networks. In: Advances in neural information processing systems. Vol. 32. 2019, p. 9240.

[56] Agarwal C, Lakkaraju H, Zitnik M. Towards a unified framework for fair and stable graph representation learning. In: Proceedings of conference on uncertainty in artificial intelligence. 2021.

[57] Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the international conference on human computer interaction. 2018, p. 1–18.

[58] Wang D, Yang Q, Abdul A, Lim BY. Designing theory-driven user-centric explainable AI. In: Proceedings of the international conference on human computer interaction. 2019, p. 1–15.

[59] Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the international conference on human computer interaction. 2020, p. 1–15.

[60] Holm EA. In defense of the black box. Science 2019;364(6435):26–7.

[61] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25(6):954–61.

[62] Kleppe A, Skrede O-J, De Raedt S, Liestøl K, Kerr DJ, Danielsen HE. Designing deep learning studies in cancer diagnostics. Nat Rev Cancer 2021;21(3):199–211.

[63] Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science 2021;373(6552):284–6.

[64] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the international conference on fairness, accountability, and transparency. 2020, p. 33–44.

[65] Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ 2020;370.

[66] Gysi DM, Do Valle Í, Zitnik M, Ameli A, Gan X, Varol O, et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. Proc Natl Acad Sci 2021;118(19).

[67] Zitnik M, Feldman MW, Leskovec J, et al. Evolution of resilience in protein interactomes across the tree of life. Proc Natl Acad Sci 2019;116(10):4426–33.

[68] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316(22):2402–10.

[69] Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018;2(3):158–64.

[70] Cao L. AI in combating the COVID-19 pandemic. IEEE Intell Syst 2022;37(2):3–13. http://dx.doi.org/10.1109/MIS.2022.3164313.

[71] Rudie JD, Rauschecker AM, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Subspecialty-level deep gray matter differential diagnoses with deep learning and Bayesian networks on clinical brain MRI: A pilot study. Radiol Artif Intell 2020;2(5):e190146. http://dx.doi.org/10.1148/ryai.2020190146, PMID: 33937838.