Review article

Towards Explainable and Trustworthy AI for Decision Support in Medicine: An Overview of Methods and Good Practices

Dimitrios Fotopoulos[1], Dimitrios Filos[1], Ioanna Chouvarda[1]

[1]*Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies School of Medicine, Faculty of Health Sciences, Aristotle University of Thessaloniki*

Abstract

Artificial Intelligence (AI) is defined as intelligence exhibited by machines, such as electronic computers. It can involve reasoning, problem solving, learning and knowledge representation, which are mostly in focus in the medical domain. Other forms of intelligence, including autonomous behavior, are also parts of AI. Data driven methods for decision support have been employed in the medical domain for some time. Machine learning (ML) is used for a wide range of complex tasks across many sectors of the industry. However, a broader spectrum of AI, including deep learning (DL) as well as autonomous agents, have been recently gaining more focus and have risen expectation for solving numerous problems in the medical domain. A barrier towards AI adoption, or rather a concern, is trust in AI, which is often hindered by issues like lack of understanding of a black-box model function, or lack of credibility related to reporting of results. Explainability and interpretability are prerequisites for the development of AI-based systems that are lawful, ethical and robust. In this respect, this paper presents an overview of concepts, best practices, and success stories, and opens the discussion for multidisciplinary work towards establishing trustworthy AI.

Keywords: Artificial Intelligence, Explainable Model, Decision Support

Corresponding Author:
Dimitrios Fotopoulos, Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki Greece. Thessaloniki 541 24, Greece, Tel.: 2310-999.247, E-mail: difoto@auth.gr

Introduction

Recent advances of AI and the success stories of the field in dealing with problems of great complexity (Samek et al.,2020; Weller, 2019), have renewed the urgency to consider the AI's socio-economic impact and the requirements to safely adopt AI solutions. Organizations in the public and private sector are collaborating for the development and establishment of strategies and policy frameworks towards lawful and ethical AI. In certain scenarios though, due to the inherent complexity and the nonlinearity that characterizes their interrelations, it is usual that the approach to their modeling and solution will carry a high degree of complexity. These models often make successful predictions and have high accuracy scores, but they are lacking in transparency, in a way that is difficult for an observer to discern how the input of the model affects its output. This also impedes the understanding of the decision-making process - i.e. *why* a decision has been made. Generally, we refer to this kind of systems, that we cannot observe their internal workings, as black boxes. Transparency may not always be necessary, but it is desired for high-stakes decision systems, such is the case of systems in the medical domain.

As AI application comes to practice in various fields, even critical ones, the matter of ethical and trustworthy AI (European Commission, 2019) has been raised. The concept ‘ethical’ means to ensure that the system does not induce harm, is fair, and its decisions can be explained/understood, and the concept ‘trustworthy’ that the system is lawful, ethical, and robust. Due to general concern of how ethical and trustworthy AI can be realized, the field of Explainable AI has emerged. Gunning and Aha define it as a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificial intelligent partners (Gunning and Aha, 2019).

Background Concepts

McCarthy, the father of Artificial Intelligence, refers to AI as the "*science and engineering of making intelligent machines*". Another definition, given from the European Commission in a document addressed to the European Parliament (2020), expands on the previous definition and explains the term by specifying how systems display intelligence - i.e. "*by analyzing their environment and taking actions -with some degree of autonomy- to achieve specific goals*".

There are numerous definitions of AI, apart from the previous ones, considering that there are many interpretations for *intelligence*, while lacking a general consensus about a single definition. A subfield of Artificial Intelligence is Machine Learning defined by Tom M. Mitchell as the study of algorithms that allow computer programs to improve through experience. ML aims to make informed decisions/predictions by generalizing from data, without the intervention of a human. Often the two terms are used interchangeably in the industry, though this is a common misconception. ML is a technique through which AI can be realized. The advancements in technology that took place in the last decade, specifically the increased computation capabilities and the vast amount of data available, transformed machine learning, and the field of AI in general, into a invaluable tool, tasked to deal with complex problems. The algorithms and models developed for this purpose, though adequately accurate to be deployed in real-world

scenarios, are not trusted yet in critical situations that involve medical diagnoses for example. A common barrier preventing this, is the uncertainty they evoke, regarding the process they follow to produce their results. Usually referred to, in the literature, as black-boxes, these AI models are required to be more transparent and explainable to be trustworthy (European Commission, 2019; Thiebes et al., 2020; Wing, 2020) The meaning of *trustworthy* can vary from one domain to another. Regarding the field of health and medicine, it may be considered as an umbrella term, which encapsulates several properties an AI model/system should incorporate, such as accountability, fairness, transparency, and privacy - among others (European Commission, 2019) . While this list is not exhaustive, and all these properties are of equal importance and interconnected with each other, we will address the issue of transparency.

Transparency can be interpreted informally as the opposite of opaque (Lipton, 2017), which is an undesired characteristic of AI black-box models. The lack of transparency, as a direct result from the increased complexity of AI models, hinders the understanding of the decision-making process, leading to AI systems that are not trusted and consequently are not adopted by the industry (Lin,2020; Ribeiro et al., 2016; The Royal Society, 2019). Transparency can have various kinds of interpretations, depending on the context and which type of user is intended for (Weller, 2019). In his article about algorithmic decision making, Diakopoulos writes that transparency can be inspected under five dimensions, namely: human involvement, data, the model, inferencing and algorithmic presence (Diakopoulos, 2016). In its general form though, transparency is aimed towards the creation of a more explainable model or system (Waltl and Vogl, 2018; Bücker et al., 2021; (European Commission, 2019). Transparency is closely related to interpretability and explainability and often used as synonyms (Doran et al., 2017), but it is highlighted through literature (Lipton, 2017; Rudin, 2019, Arya et al., 2019) that they convey a different meaning, so in this paper they will be used distinguishably.

Interpretability addresses the issue of conveying some of the properties of an ML model in terms understandable to a human (Roscher et al., 2020). It aims to increase the understandability of the system, which has been proposed in relevant work to be used as a evaluation metric (Allahyari and Lavesson, 2011), by clarifying the link between the prediction of the machine learning model and its selected features. It is worth mentioning that with understandability we refer to the functional aspects of the system, and not to its technical inner mechanisms, as it is also mentioned in (Lipton, 2017). Explainability, on the other hand, does not have a clear definition, although it has been recognized that is an important characteristic of AI models (Roscher et al., 2020). It is described in relevant literature as, the knowledge of what each component of the system represents and its contribution towards the system's results. In the document containing guidelines for trustworthy AI, from European Commission (2019), it is also stated that explainability is about the technical processes of the model and the related human decisions made by it. Holzinger (Holzinger et al., 2017) does not provide a distinct definition for explainability; instead, the author explains that there are two types of interpretability/explainability - namely one that explains what is already interpretable (post-hoc explainability) and one that builds explainable components into the

structure of an AI model (ante-hoc explainability).

An enhancement to the interpretability/explainability endeavor of black box models is to try and give insights about the causal relationships that are shaped. As previously mentioned (Doshi-Velez and Kim, 2017), interpretable ML, *causality implies that the predicted change in output due to a perturbation will occur in the real system*, Approaches that address causality in the, can improve the interpretable characteristics of a model by providing information about the contribution of its components in its final decisions. Moraffah et al. distinguishes between traditional interpretability, or else *statistical*, and causal interpretability, which aims to answer questions of the type "What-if" (causal interventional interpretability) and "Why" (counterfactual interpretability) (Moraffah et al., 2020). ML models are capable of discerning associations and correlations in a vast amount of data, but they cannot provide causal explanations for these. A relevant term is causabilty, as Holzinger defines it (Holzinger et al., 2019), distinguishing it from causality, as a property which examines causality from the user scope and is measured for how understandable and transparent is to a human expert. Defining cause-effect relations is a method to deal with data bias in models and make them more robust to it.

Additionally, when considering interpretability, bias is a major topic to the field. All data-driven methods are expected to be built upon diverse datasets, that represent the actual diverse patient populations it addresses for diagnosis and treatment support. Data selection bias occurs to some extent with any data set, due to its limited volume. This bias often originates from over-representation or under-representation of groups or subsets based on gender, ethnic, social, environmental, or economic factors, as well as health-related confounding factors, like comorbidities or treatments. Sometimes, bias is introduced by technical or organisational processes, like methods of labeling, post-processing, and annotating (Geis et al., 2019).

Existing Tools and Approaches

The current report will focus on the aspect of the transparency of the model (not the algorithm), a concept that serves to enhance its explainability. It can be achieved through two approaches or their combination in some cases (Molnar, 2019): the deployment of interpretable models - models that are sufficiently transparent by their nature and thus understandable to the user, or the utilization of explanation methods, that are applied after the deployment of a model.

In the first category fall under, according to recent literature (Freitas, 2014; European Commision, 2020; Thiebes et al., 2020; Huysmans et al., 2011) these models: decision tree, rules, linear models. Linear regression models are used widely in various fields due to their inherent interpretability. This is based on their ability to quantify the outcome of a prediction and by showing the degree of influence of each variable. Linear models make predictions based on the weighted sum of the features of a model's instance, hence their interpretability lies in the interpretation of their features. This can be done by considering the feature importance, which is basically the contribution of each feature for a given prediction. This becomes evident by visualizing them. There is an option of visualizing the various features' weights (*weight plot*), but this has the disadvantage of measurements being on a different scale. As an alternative solution, we can plot the weights multiplied by their

feature (*effect plot*), in order to understand how much the features and their weights influence the outcome. The greatest advantage of linear models, which is their linearity, is also one of its limiting factors. Because of their simplistic nature, they cannot model sufficiently the complexity of real-world scenarios. Furthermore, any interaction between features that creates nonlinearities must be addressed by forming new input features based on knowledge (Molnar, 2019).

Decision rules, most often of the form IF-THEN, are also one of the intrinsically interpretable models. A decision rule is a function which maps an observation to an appropriate action. The *if* clause is a condition or a combination of conditions conjuncted with the logical *AND*, and the *then* part is the prediction. Decision trees have a graph structure that resembles a tree. A decision tree model starts from a node (root) that branches into other nodes that represent a question (test) for a feature, and they also branch to a child node for each possible answer or to another node with a different input feature. The path from the root to the leaves of the tree represent the rules that the classification is based upon and they can be extracted. Trees have the advantage of being able to deal with categorical and continuous variables, in contrast with decision rules, where the features have to be of the former type (Kingsford and Salzberg, 2008). Interpretations can be extracted by examining the structure of decision tree and rules models and tracing how they make their predictions (Molnar et al., 2020). This however becomes quite challenging with complex scenarios, in which they require a large number of rules and features that interact with each other or a high degree of depth in the decision tree. Both decision trees and rules are quite interpretable, although their representation of interpretation varies. Decision trees have a strong visual characteristic due to their graphical representation, whereas decision rules interpretation is based on their textual form (Guidotti et al., 2019).

These models are inherently interpretable or, as often mentioned in the literature, *white box* models (Reyes et al., 2020). In contrast, when ML models do not have interpretable properties, certain techniques are utilized for extracting explanations after the deployment of the model. These methods that operate on black box models are post-hoc explainability methods (Holzinger et al., 2017). In this work, we will focus on methods that are not model specific. According to the taxonomy proposed by Arrieta et al, *model-agnostic* techniques approach explainability through model simplification, feature relevance estimation and visualization techniques (Arrieta et al., 2020). A common technique for model simplification is *Local Interpretable Model-Agnostic Explanations (LIME)* (Thiebes et al., 2020). LIME attempts to formulate explanations for the predictions of a model, by approximating it with a more simple, interpretable model. It focuses on local explainability, i.e., attempts to explain individual predictions of the model. Ribeiro et al. report that by using LIME, even non-expert users can be benefited from provided explanations, something that may contribute towards scenarios where a trustworthy AI model is of foremost importance.

Another group of techniques to generate explanations for black box models is visualization. Visualization tools and techniques can be powerful methods to convey information in a human-interpretable manner. These post-hoc visual explanations aim to interpret the model's behavior through visual components. Representative examples of this category of methods are *individual conditional expectation*

(ICE) (Goldstein et al., 2014) and *partial dependence plots (PDP) (*Friedman, 2001). The purpose of PDP is to offer insights about the importance of some of the model's features on the predicted outcome. According to Friedman, it can be useful to depict their relationship and help with the interpretability of a model, but its result can be misleading for highly correlated features. While PDP focuses on the global effect a feature has on a prediction, ICE highlights the influence of an individual contribution (Goldstein et al., 2014). Both methods apply the concept that if you make alterations to a value of an important feature, then it is expected to show at the model's outcome. Reyes et al. argue that PDP and ICE could prove useful in the field of radiology, since features are not generated by an algorithm but are hand-crafted based on prior knowledge, which can be validated through the former visualization techniques (Reyes et al., 2020).

For techniques that produce explanations by examining the influence of each feature, a characteristic example may be *Shapley additive explanations* (*SHAP)* (Lundberg and Lee, 2017), based on the concept of Shapley values from game theory. In this analogy, the game is the result decision of the model and the players are the features. According to Lundberg and Lee, SHAP is a unified framework of other explainability methods. It aims to interpret the prediction of an instance of the machine learning model (game) by measuring the contribution of its features (players) to that prediction. From the definition of the method, it is evident that it focuses on local interpretability. But it is feasible to provide global interpretations for the model, by aggregating the required Shapley values for each feature.

Exemplary Applications in the Health Domain

One of the main barriers regarding the application of AI tools in medicine is the inability of the medical professionals to understand the rationale behind specific decisions proposed by the algorithms and thus, XAI seems to be vital for the integration of AI into decision support systems.

One of the fields that a significant effort was paid is the management of critically ill patients. The medical condition of such patients may change dramatically in time while a plethora of heterogeneous data are available for each of them. In this respect, accurate decision making is crucial, while time plays a vital role. Most XAI methods that have been implemented for ICU patients focus on highlighting indented feature importance. According to Ge et al., this approach was followed in order to identify the most contributing features for the prediction of mortality in ICU patients (Ge et al., 2018). On the other hand, Kaji et al. focused on the identification of those parameters that could predict the initiation of critical events during patients' stay in ICU (Kaji et al., 2019), while Shickel et al. proposed a score that could accurately predict patient's severity of illness during an ICU stay (Shickel et al., 2019).

The outbreak of the COVID-19 pandemic affected the research related to XAI. In this respect, a significant effort was paid to implement AI models which can quantify the disease, stratify patients, and predict outcome. Most works use XAI to provide more information regarding the disease detection using CT scans or chest radiograph (X-ray) (Ahsan et al., 2020). XAI is succeeded through the provision of powerful visualization and confidence scores for each layer of the DL model. A characteristic recent work (Chassagnon et al., 2021) proposes the

use of imaging characteristics extracted from CT scans (radiomics), while they combine them with diverse types of data, such as clinical and biological attributes to select the most significant ones, regarding the outcome prediction (severe and non-severe cases). The incorporation of visualization aspects allows the implementation of a COVID-19 Holistic Multi-Omics Signature & Staging mechanism, leading to data augmentation and improving the explainability of the predictive models.

Discussion: Technical and Methodological Challenges

As machine learning is increasingly used in real-world decision processes, the necessity for transparency will continue to grow. The emerging field of explainable AI in the medical domain holds promising results towards more transparent machine learning models and a broader adoption in the healthcare. In this direction, there are several challenges to overcome, including:

1. what is an explanation, and how it depends on the problem (i.e. a diagnostic case) and the user (i.e. clinical expert or other user)
2. what are the metrics for comprehensibility, and how can they be contextualized

Many questions still remain to be solved, towards creating a formal framework in explainable AI, with extensions to trustful AI. Such a framework is foreseen to form a general background regarding concepts that are not well defined yet and create a common taxonomy , in order to promote further research and facilitate comparison between related works. In addition, a XAI framework is expected to formalize the tools and best practices thar can boost explainability in different scenarios and contexts.

Conclusion

This paper serves as an introduction of the field of explainable AI and comes as an effort to lay out the outline, without overwhelming the reader. We reviewed the concept of what trustworthy AI is and what are the challenges that come from using black box AI models. We approached the definition of trustworthy AI from the domain of transparency, as this quality is often used to confirm other desired aspects of AI models.

References

Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M. L., & Shakhawat Hossain, M. (2020). COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. Machine Learning and Knowledge Extraction, 2(4), 490–504. https://doi.org/10.3390/make2040027

Allahyari, H., & Lavesson, N. (2011). User-oriented Assessment of Classification Model Understandability. In: Electronic Research Archive of Blekinge Institute of Technology, 11th Scandinavian Conference on Artificial Intelligence. Trondheim, Norway, 24-26 May 2011. IOS Press

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2019). One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. ArXiv:1909.03012 [Cs, Stat]. http://arxiv.org/abs/1909.03012

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bücker, M., Szepannek, G., Gosiewska, A., Biecek, P. (2020). Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring. ArXiv:2009.13384 [stat.ML]. https://arxiv.org/abs/2009.13384

Chassagnon, G., Vakalopoulou, M., Battistella, E., Christodoulidis, S., Hoang-Thi, T.-N., Dangeard, S., Deutsch, E., Andre, F., Guillo, E., Halm, N., El Hajj, S., Bompard, F., Neveu, S., Hani, C., Saab, I., Campredon, A., Koulakian, H., Bennani, S., Freche, G., … Paragios, N. (2021). AI-driven quantification, staging and outcome prediction of COVID-19 pneumonia. Medical Image Analysis, 67, 101860. https://doi.org/10.1016/j.media.2020.101860

Diakopoulos, N. (2016). Accountability in Algorithmic Decision Making. Available at: <https://cacm.acm.org/magazines/2016/2/197421-accountability-in-algorithmic-decision-making/fulltext>. [Accessed 31 January 2021]

Doran, D., Schulz, S., & Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. ArXiv:1710.00794 [Cs]. http://arxiv.org/abs/1710.00794

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. ArXiv:1702.08608 [Cs, Stat]. http://arxiv.org/abs/1702.08608

European Commision. (2018). Communication Artificial Intelligence for Europe. Shaping Europe's Digital Future. Available at: <https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe> [Accessed 31 January 2021]

Freitas, A. A. (2014). Comprehensible classification models: a position paper. ACM SIGKDD Explorations Newsletter, 15(1), 1–10. https://doi.org/10.1145/2594473.2594475

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5):1189–1232

Ge, W., Huh, J.-W., Park, Y. R., Lee, J.-H., Kim, Y.-H., & Turchin, A. (2018). An Interpretable ICU Mortality Prediction Model Based on Logistic Regression and Recurrent Neural Networks with LSTM units. AMIA Annual Symposium Proceedings, 2018, 460–469. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371274/

Geis, J. R., Brady, A., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., Langer, S. G., Kitts, A. B., Birch, J., Shields, W. F., van den Hoven van Genderen, R., Kotter, E., Gichoya, J. W., Cook, T. S., Morgan, M. B., Tang, A., Safdar, N. M., & Kohli, M. (2019). Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. Insights into Imaging, 10(1), 101. https://doi.org/10.1186/s13244-019-0785-8

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2014). Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. ArXiv:1309.6392 [Stat]. http://arxiv.org/abs/1309.6392

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. and Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program, AI Magazine, 40(2), pp. 44-58. doi: 10.1609/aimag.v40i2.2850.

European Commission. (2019). Ethics Guidelines for Trustworthy AI, Technical Report, . Available at: <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai> [Accessed 31 January 2021].

European Commission.(2020). White Paper on Artificial Intelligence: Public consultation towards a European approach for excellence and trust, European Commission. Available at: <https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf> [Accessed 31 January 2021]

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? ArXiv:1712.09923 [Cs, Stat]. http://arxiv.org/abs/1712.09923

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery, 9(4). https://doi.org/10.1002/widm.1312

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. Decision Support Systems, 51(1), 141–154.

https://doi.org/10.1016/j.dss.2010.12.003
jfr-paragios.pdf. (n.d.).

Kaji, D. A., Zech, J. R., Kim, J. S., Cho, S. K., Dangayach, N. S., Costa, A. B., & Oermann, E. K. (2019). An attention based deep learning model of clinical events in the intensive care unit. PloS One, 14(2), e0211057. https://doi.org/10.1371/journal.pone.0211057

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? Nature Biotechnology, 26(9), 1011–1013. https://doi.org/10.1038/nbt0908-1011

Lin, Z. Q. (n.d.). Quantifying the Performance of Explainability Algorithms. 72.

Lipton, Z. C. (2017). The Mythos of Model Interpretability. ArXiv:1606.03490 [Cs, Stat]. http://arxiv.org/abs/1606.03490

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. ArXiv:1705.07874 [Cs, Stat]. http://arxiv.org/abs/1705.07874

Merrick, L., & Taly, A. (2020). The Explanation Game: Explaining Machine Learning Models Using Shapley Values. ArXiv:1909.08128 [Cs, Stat]. http://arxiv.org/abs/1909.08128

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. ArXiv:2010.09337 [Cs, Stat]. http://arxiv.org/abs/2010.09337

Molnar, C., 2019. Interpretable machine learning. A Guide for Making Black Box Models Explainable, Available through: <https://christophm.github.io/interpretable-ml-book/> [Accessed 31 January 2021]

Moraffah, R., Karami, M., Guo, R., Raglin, A., & Liu, H. (2020). Causal Interpretability for Machine Learning - Problems, Methods and Evaluation. ACM SIGKDD Explorations Newsletter, 22(1), 18–33. https://doi.org/10.1145/3400051.3400058

Observatory of Public Sector Innovation. (2019). AI Strategies & Public Sector Components. Retrieved January 31, 2021, from https://oecd-opsi.org/projects/ai/strategies/

Reyes, M., Meier, R., Pereira, S., Silva, C. A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R. M., & Wiest, R. (2020). On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. Radiology. Artificial Intelligence, 2(3), e190043. https://doi.org/10.1148/ryai.2020190043

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access, 8, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199

Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. ArXiv:1811.10154 [Cs, Stat]. http://arxiv.org/abs/1811.10154

Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2020). Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. ArXiv:2003.07631 [Cs, Stat]. http://arxiv.org/abs/2003.07631

Shickel, B., Loftus, T. J., Adhikari, L., Ozrazgat-Baslanti, T., Bihorac, A., & Rashidi, P. (2019). DeepSOFA: A Continuous Acuity Score for Critically Ill Patients using Clinically Interpretable Deep Learning. Scientific Reports, 9(1), 1879. https://doi.org/10.1038/s41598-019-38491-0

The Royal Society. (2019). Explainable AI: the basics. Available at: <https://ec.europa.eu/futurium/en/system/files/ged/ai-and-interpretability-policy-briefing_creative_commons.pdf> [Accessed at 31 January 2021]

Thiebes, S., Lins, S. & Sunyaev, A. (2020). Trustworthy artificial intelligence. Electron Markets. https://doi.org/10.1007/s12525-020-00441-4

Tjoa, E., & Guan, C. (2020). A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 1–21. https://doi.org/10.1109/TNNLS.2020.3027314

Waltl, B., & Vogl, R. (2018). Increasing Transparency in Algorithmic-Decision-Making with Explainable AI. Datenschutz Und Datensicherheit - DuD, 42(10), 613–617. https://doi.org/10.1007/s11623-018-1011-4

Weller, A. (2019). Transparency: Motivations and Challenges. ArXiv:1708.01870 [Cs]. http://arxiv.org/abs/1708.01870