Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

© 2022 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/SHTI210854

Medical Informatics in a Tension Between Black-Box AI and Trust

Murat SARIYAR[a,1] and Jürgen HOLM[a]
[a]*Bern University of Appl. Sciences, Department of Medical Informatics, Switzerland*

Abstract. For medical informaticians, it became more and more crucial to assess the benefits and disadvantages of AI-based solutions as promising alternatives for many traditional tools. Besides quantitative criteria such as accuracy and processing time, healthcare providers are often interested in qualitative explanations of the solutions. Explainable AI provides methods and tools, which are interpretable enough that it affords different stakeholders a qualitative understanding of its solutions. Its main purpose is to provide insights into the black-box mechanism of machine learning programs. Our goal here is to advance the problem of qualitatively assessing AI from the perspective of medical informaticians by providing insights into the central notions, namely: explainability, interpretability, understanding, trust, and confidence.

Keywords. Artificial intelligence, Explainable AI, Luhmann, Confidence, Trust

1. Introduction

Artificial intelligence (AI) is now essential part of many activities in the field of medical informatics, not only in research but also in the healthcare setting [1, 2]. The first FDA-approved medical device that relies on AI was the BodyGuardian ® Remote Monitoring System from Preventice Solutions in 2012, which detects cardiac rhythm abnormalities, using small wearable monitors paired with a dedicated smart-phone [3]. Main reason for the success story of AI is the boost in prediction accuracy due to advances in digital documentation, computing power, (deep learning) algorithms, and wearable medical devices. Relevant application areas are disease diagnosis, image classification, natural language processing of electronic health records, biomarker discovery, and drug development.

Even for medical informaticians not developing AI-based decision support systems, it became crucial to assess the benefits and disadvantages of AI-based solutions as promising alternatives for many traditional tools. Besides quantitative criteria such as accuracy and processing time, healthcare providers are often interested in qualitative explanations of the solutions. For the former issue, medical informaticians could rely more and more on the results of explainable AI (XAI) approaches [4]. However, there are still many trust problems. As all three terms “explainability”, “responsibility”, and “trust” are often not clarified, qualitative assessments of AI frequently are frequently not satisfying, independently of relying on XAI or not.

[1] Corresponding Author, Murat Sariyar, Bern University of Applied Sciences, Quellgasse 21, CH2502 Biel/Bienne, Switzerland; E-mail: murat.sariyar@bfh.ch.

Our goal here is to advance the problem of qualitatively assessing AI from the perspective of medical informaticians by providing insights into the central notions and their relations. As a use case, we will refer to biomarker discovery for pharmacogenes as a step towards a comprehensive decision support systems concerning drug therapies. Pharmacogenetics investigates the association between genetic variations and the drug response to help tailoring pharmacotherapy to the individual patients' characteristics, thereby avoiding unnecessary adverse drug events (ADE) and increasing therapeutic efficacy [5]. Especially stratification according to the origin of individuals is an urgent issue, which requires augmenting the current biomarkers with the help of AI.

2. Methods

A first step for the qualitative assessment of AI is the definition of AI as well as XAI and the listing of approaches for XAI relevant for making the distinction between explanation and interpretation. After that, the term explanation is defined with reference to scientific theory. Understanding, trust, and confidence are defined through reference to system theory of Luhmann, as it provides a holistic foundation of these notions, which allows to detect discrepancies in their empirical use [6,7]. The impacts of these definitions are discussed with respect to our use case.

3. Results

AI is the "theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages" [8]. Usually weak AI is assumed: computer systems are made to act as if they were intelligent, not to be intelligent. Two high-level approaches for achieving such intelligence are rule-based and non-rule-based. The latter is associated with machine learning, i.e., a computer program learns it tasks from available data and improves its performance with further data from the same context. AI methods are used as a basis for many clinical expert systems, which suggest solutions to problems that were previously solved by human experts alone. Such a genealogy and the relevance of medical decisions raise the demand for interpretations of AI solutions.

XAI provides methods and tools, which are interpretable enough that it affords different stakeholders a qualitative understanding of its solutions [9]. Its main purpose is to provide insights into the black-box mechanism of machine learning programs. The term "black-box" refers to the opaqueness related to the mechanisms responsible for producing solutions. Approaches for opening the black-box can be differentiated into model-agnostic and model-specific. Examples for the former ones are sensitivity analysis and local interpretable models (LIME). Both are not looking under hood, but tweak the input and observe the resulting effects, thereby gaining insights into the relevance of features. LIME is often used in practice, as it can model interactions by fitting surrogate linear models to the results of multifeatured perturbations around certain predictions (hence local). Model-specific XAI methods rely either on the mechanism of the algorithm itself – e.g., decision trees allow to trace the decision paths and to extract the key determinants – or provide means for looking under the hood in a certain class of methods, e.g., relevance propagation in the case of deep neural networks.

What should be achieved by XAI? Explainability as the name indicates, interpretability as the definition above suggests or understandability? Is the final goal trust or confidence? We locate many acceptance problems of AI in the lack of clarity and relevance assessment with respect to these notions. Explainability refers to the possibility of providing reasons (explanans) for the outcome (explanandum) of a system. There are many different forms for explanation: the deductive-nomological, inductive-statistical, causal mechanical, etc. [10]. In all cases, it is central to provide justification (reasons) that are comprehensible and transparent (details) at the appropriate level for the addressed audience. For example, in terms of causal mechanical explanation, an explanandum X explains an explanans C, if (i) X increases the probability of C, given the other explanatory factors F (statistical relevance $P(C \mid F \,\&\, X) > P(C \mid F)$), and (ii) X fits into the causal nexus of the explanans C. An explanation for "Why does this *deep neural net* provide the *best solutions*?" would generally refer to components of the deep neural net at different levels that increases the probability of making the right predictions. The problem is that they are multiple explanations, and most of them won't be satisfying for a non-expert. Hence, pure explainability is not sufficient for the goals of XAI.

Interpretability is a property of an explanation that describes the extent to which the cognitive capacities of the addressed audience is taken into consideration. Hence, this notion highlights, that explanation is a social process for which understanding of the addressed audience should come into play. According to Luhmann, understanding is the result of constructing a distinction between the information provided and its form (as a text or verbally), which is only retrospectively perceivable in following communication events [11]. In other words, understanding is only measurable through ensuing activities and utterances of the addressed audience. The advantage of this definition is, that it can ignore the unsolvable problem of how to avoid the case of where someone can explain something without having understood it. If the following communication is compatible with the goal of XAI, we are fine. This does not provide a solution to the appropriate level of explanations; it just indicates the necessity for an – often iterative – social process for achieving satisfying interpretability.

The most important result of understanding in the context of XAI should be either trust or confidence. Following Luhmann again, trust is a decision to rely on one's own expectations with respect to certain mechanisms in view of the involved risks and alternatives. Confidence on the other hand is a reliance on one's own expectations concerning mechanisms without taking alternatives or the risks into consideration. In both cases, complexity is reduced by taken something for granted. Within the context of AI, confidence is sufficient, if the decision-making system is related to non-sensitive data, for example in the context of search engines or recommender systems. Especially, for clinical AI application, it can be important to arrive at trust. Health care providers know at least one alternative to an AI system: the human decision making. One should distinguish two levels of trust: layman trust, for which high-level explanations are sufficient and expert trust, which require many details. Table1 summarizes the insights concerning XAI in terms of central properties.

Table 1. Properties of different AI goals with one example for each of the goals.

Goals of XAI	Black box	Alternatives	Examples
Confidence	yes	Not Considered	"The system is already in use in hospital X"
Layman trust	Opened slightly	Considered	List of features that have significant impacts (LIME)
Expert trust	Opened fully	Considered fully	Results of relevance propagation for deep nets

For our use case, confidence means that a new pharmacogenetic biomarker signature for a certain subgroup produced by an AI algorithm will be accepted by physicians not familiar with biomarker discovery if, for example, working groups such as CPIC (Clinical Pharmacogenetics Implementation Consortium) will validate it. Layman trust is necessary, if other biomarkers for the subgroup are known to be inferred through classical statistical methods, in which case we provide high-level insights for physicians through LIME. Expert trust can be required for the CPIC validation, which must ensure that there are no biases in using certain AI implementations.

4. Discussion

An implication of our results is the requirement for more stakeholder involvement, especially in the case of translational research. There is no one-fits-all solution of XAI. It needs to be adapted to the context and via a (n iterative) social process, which means to augment the available methods by diversified qualitative explanations. We are aware of the fact, that scientists are not eager to invest time in such processes, but instead of regulating it or creating working groups for discussing how to foster such a culture, there should be an intrinsic motivation for appropriate explanations. Physicians can provide much better feedback if they understand mechanism behind solutions.

Further research in qualitative XAI should focus on informing quantitative XAI methods and vice versa. A comprehensive categorization of available XAI methods in terms of their usefulness for qualitative explanations would also be useful for these methods themselves, as this could foster an understanding for which audiences these methods are developed and how they should be improved. In addition to that, system theory provides a rich arsenal of explanations regarding social settings of trust and understanding, which should be referred to more often. We are confident that AI will gain more trust through adapted social practices, not by ex-cathedra statements.

References

[1] Emanuel EJ, Wachter RM. Artificial Intelligence in Health Care: Will the Value Match the Hype? JAMA 321 (2019), 2281–2282.
[2] Choudhury A, Asan O. Role of Artificial Intelligence in Patient Safety Outcomes: Systematic Literature Review. JMIR Med Inform 8 (2020), e18599.
[3] Teplitzky BA, McRoberts M, Ghanbari H. Deep learning for comprehensive ECG annotation. Heart Rhythm 17 (2020), 881–888.
[4] Holzinger A, Langs G, Denk H, et al. Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip Rev Data Min Knowl Discov 9 (2019), e1312.
[5] Wake DT, Ilbawi N, Dunnenberger HM, et al. Pharmacogenomics: Prescribing Precisely. Med Clin North Am 103 (2019), 977–990.
[6] Luhmann N. Trust and Power. Chichester: Wiley, 2017.
[7] Pieters W. Explanation and trust: what to tell the user in security and AI? Ethics Inf Technol 13 (2011), 53–64.
[8] Walsh K. Artificial intelligence and healthcare professional education: superhuman resources for health? Postgrad Med J 96 (2020), 121–122.
[9] Duell J, Fan X, Burnett B, et al. A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 2021, pp. 1–4.
[10] Dammann O. Evidence Mapping to Justify Health Interventions. Perspect Biol Med 64 (2021), 155–172.
[11] Luhmann N. Social Systems. Stanford: Stanford University Press, 1996.