Going Beyond Explainability in Medical AI Systems

Ulrich Reimer[1] Edith Maier[1] Beat Tödtli[1]

Abstract:

Despite the promises of artificial intelligence (AI) technology, its adoption in medicine has met with formidable obstacles due to the inherent opaqueness of the internal decision processes that are based on models which are difficult or even impossible to understand. In particular, the increasing usage of (deep) neural nets and the resulting black-box algorithms has led to wide-spread demands for explainability. Apart from discussing how explainability might be achieved, the paper also looks at other approaches at building trust such as certification or controlling bias. Trust is a crucial prerequisite for the use and acceptance of AI systems in the medical domain.

Keywords: Explainability; Certification; Machine Learning; AI; Bias; Medical Device

1 Introduction: Medical AI Systems

AI systems are beginning to have an impact in the medical domain. They comprise applications for clinicians that allow rapid and accurate image interpretation, genome sequencing or recommend treatments for patients (such as IBM Watson Health's cancer algorithm). AI technology can also be found in applications for lay people such as biosensors with continuous output of physiologic metrics that enable them to process their own data to promote their health. According to [9], for these purposes, algorithms are desperately needed because analysing such huge amounts of data exceeds human capacities. The question is how AI and human intelligence can best be integrated.

The decision on a patient's next treatment is shared between the patient, the physician and the medical AI system[2]. The degree of autonomy and responsibility in the decision-making process of each role is central to the attribution of trust since clearly, systems require less trust if they are supervised by a human or another system.

Therefore, it may be useful to distinguish between decision *support* systems and decision *making* systems, as well as recognising the continuum between these extremes. For example, a skin melanoma detection system may be viewed as a decision making system if false-negative detections are not re-examined by a physician. Medical information retrieval systems on the other hand are clearly only capable of providing decision support.

[1] Institute for Information & Process Management, University of Applied Sciences St.Gallen, firstname.lastname@fhsg.ch

[2] One may object to the anthropomorphic notion of a system making a decision. Alternatively, the decision of an AI system may be viewed as a shared decision by the physician and the engineers who built the system.

2 Trust-Enhancing Features of Medical AI Systems

An analysis of trust-influencing features of automated systems in general is given by Chien et al. [1]. The authors suggest three main categories, each having several dimensions some of which are particularly relevant for our topic:

- The category Performance Expectancy is defined "as an individual's belief that applying automation will help her to enhance job performance". One particularly relevant dimension of this category is *perceived usefulness*, which we will discuss briefly in Section 2.5.
- The second category Process Transparency refers to factors that influence an individual's perceived difficulty in using an automated system. Relevant dimensions are *understandability* (see Sec.2.1) and *reliability* (see Sec.2.4) of the system.
- Finally, the category of Purpose Influence relates to "a person's knowledge of what the automation is supposed to do", i.e. the alignment of a user's expectations and the system designers' goals. An important dimension of this category is the *certification* of an automated system, which will be discussed in Section 2.2.

An additional aspect, specific to data-driven AI systems and not mentioned in [1] is controlling the sampling bias. We will deal with this aspect in Section 2.3.

In this position paper we focus on the following questions:

- What makes a user trust a (medical) AI system to provide correct and adequate advice or decisions?
- Will (medical) AI systems override experience and intuition, i.e. the largely tacit knowledge harboured by experts, when it comes to taking decisions?

While the above mentioned issues are relevant for AI systems in general, they deserve particular attention in medicine where lives may be at stake. The models AI systems are based on primarily determine their behaviour. We therefore argue that certain characteristics of these models such as *interpretability* and *representativeness* play a crucial role for the usage and acceptance of AI systems.

In the following subsections we will have a closer look at each of the above mentioned dimensions and discuss how they may influence a physician's or patient's trust in a medical AI system.

2.1 Explainability

The trust-enhancing dimension of understandability is closely related to the notion of *explainability* as discussed in the machine learning community [2]. Gilpin et al. [3]

distinguish between explainability and interpretability. According to them the goal of interpretability is to describe the internals of a system in a way that is understandable to humans. Since humans are biased towards simple descriptions, an oversimplification of the description of the actual system is unavoidable and even necessary. Explainability includes interpretability but additionally requires the explanation to be *complete*, i.e. not lacking relevant aspects. Clearly, there is a trade-off between an explanation being complete and comprehensible. Other authors refuse to make a distinction between explainability and interpretability (e.g. [6]).

Explainability is typically problematic for sub-symbolic models, i.e. (deep) neural networks as opposed to symbolic models such as decision trees, which can in principle be inspected by a human. Nevertheless, inspecting a complex decision tree with hundreds or even thousands of nodes would quite probably be pointless since reading does not automatically imply understanding. Thus we run into the *performance-explainability-tradeoff* between having simple (symbolic) models that facilitate explainability and complex (possibly sub-symbolic) models that result in a better performance of the AI application but cannot be understood anymore [5].

Since sub-symbolic models cannot (easily) be inspected, several researchers have come up with the idea of having an additional, simpler (symbolic) model just for the purpose of explainability while the actual system performance relies on the full-fledged, complex (sub-symbolic) model [7].

Other authors suggest to transfer a part of the input-to-output transformation complexity from the modelling to the preprocessing stage [8]. This may help to improve the explainability of a model at the expense of a more complex and opaque preprocessing procedure .

It can be stipulated that explainability or at least interpretability is an essential capability of a medical AI system. Otherwise, a physician or clinician either just has to trust the conclusions of the system or has to go through a subsequent verification process, which may well be costly and time-consuming and thus nullify any potential efficiency benefits of the AI system. At the same time, he or she may not be willing to go through a lengthy explanation to understand the decision offered by the system. Since neither approach is desirable or practicable, the task of verification is normally taken on by officially recognised agencies or regulatory bodies such as the Food and Drug Administration (FDA).

2.2 Certification of Medical AI Systems

Medical AI systems need to get certified as medical devices by regulatory bodies such as the already-mentioned FDA in the US or national bodies such as SwissMedics in Switzerland before they can be used in practice. A *clinical trial* is usually at the core of the certification process. By assuming responsibility for the adequacy of the medical AI system, regulatory bodies provide an established source of trust.

However, the certification of a medical AI system requires a different approach than that for approving e.g. a new drug. The decisions suggested by an AI system must be compared to the decisions of a physician, whereas clinical trials evaluate new treatments by comparing them with traditional or treatment as usual. Since the model underlying an AI system is a generalization of a possibly large but limited input it will sometimes come up with inadequate decisions. Thus, for AI systems to be certifiable at all it is important that other criteria are fulfilled – e.g. that a clinician can easily cross-check the systems' decision against his/her own expertise, by simple lab tests or the explanation given by the system.

Depending on its complexity, the certification of an AI system can require a huge effort. One possibility to simplify the process is to narrow down the *functional range* of the system, e.g. by having it diagnose only one kind of disease or a small range of diseases. The downside of breaking up the diagnostic scope of an AI system into smaller systems poses the problem that a disease might not be diagnosed if it comes with atypical symptoms. A system with a broader range can more easily do a differential diagnosis between several possible diagnoses.

Another approach to narrow down the scope of an AI system is the *range of patients* it can be applied to. For example, if it was developed on data from a group of people of a specific ethnic group and gender the clinical trial can be narrowed down to the same kind of sample of applicants. This approach would also help with the issue of bias, as is further discussed in Section 2.3.

Certification amounts to *model testing*. This means the absence of errors (wrong diagnoses, wrong therapies) cannot be shown. When the certification process uses a sample with a similar bias as the sample used for developing the AI system there might exist fundamental flaws that will not be uncovered during certification. Thus, instead of only doing model testing via a clinical trial, a certification process should additionally include inspecting the model and checking it for plausibility. Here, an explanation component could provide considerable support and help to “ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected"[3].

2.3 Reducing Sampling Bias in medical AI Systems

Sampling bias refers to the bias incurred due to the specific data set chosen to train an AI system. The resulting system extrapolates from the training data to the general case. If the training set is skewed the system does not generalize well. For example, if a medical AI system is trained on data from Asian people it might not work well for Africans or Europeans. While the bias concerning gender and ethnic group can be controlled relatively easily [10], the feature space is huge so that other, less obvious biases can exist that neither the developer nor the certification agency are aware of. The problem is that it is usually unknown what the effect of a feature on the generated model is and how its values should be distributed to provide a representative sample.

Bias can be reduced by utilizing *domain knowledge* about features and their impact on the learned model [4]. Another approach is testing for specific biases. When using medical AI systems a patient may rightfully ask how well she or he is represented in the training data. An approach to do this might be to go back to the original dataset and determine the nearest neighbours of the data point representing the patient. Comparing the number and closeness of the nearest neighbours to the number and closeness of the average person in the sample gives an estimate how well that patient is covered or if he or she is an outlier.

2.4 Reliability

The reliability of an AI system may influence the willingness to use it [1]. Reliability can be estimated by classical measures from machine learning such as precision, specificity, sensitivity (or recall), false positive rate and false negative rate. An important aspect in this context is if the system is optimized to reduce false positives or false negatives.

2.5 Perceived Usefulness

Finally, a further criterion for deciding to use a (medical) AI system is its perceived usefulness, which refers to a user's belief that the system would be of assistance to achieve a certain goal [1]. This criterion is strongly influenced by reliability (see above) but also by other factors such as the degree of support the system gives and how often it fails to comes up with a viable solution.

3 Points for discussion

Let us summarize the main points to be addressed when discussing the prerequisites for accepting AI technology in the medical domain:

- Physicians do not need an explanation component because a medical AI system has already undergone a formal certification process which relieves them from the responsibility to ensure that advice obtained from the system is correct.
- Even if a medical AI system has obtained official certification, clinicians might at least want to check the plausibility of a system's decision. Therefore an explanation component should be available that provides succinct and intuitive (but not necessarily complete) explanations.
- AI systems can only be certified if the system is not too complex, i.e. has a limited functional range and/or patient range. Otherwise the required effort would make certification too costly and impractical.

- Certification should not only focus on system performance, i.e. on testing the underlying model with concrete cases, but also include the inspection of the underlying decision model as well as the training and test sets used for creating the model.
- An AI system might be more reliable than a human expert, make decisions in less time and with lower costs and come up with better suggestions, but is still bound to make errors eventually. The error rate, however, should be inferior to that of a physician.
- The sample used to train the decision model of an AI system will always be biased. To make up for this it should be possible to determine for specific patients how well they are covered by the training sample to get an estimate how much to trust the AI system in those specific cases.
- Instead of applying a pure machine learning approach to generate the model underlying an AI system it should also comprise (manually engineered) domain knowledge to increase its reliability, e.g. for plausibility checking of a decision.

4 Conclusions and Future Work

As we have seen there are approaches to building trust in medical AI systems that go beyond explainability such as certification, controlling the bias of the training set or checking how well an individual patient is covered by the learned model. But even if all these are applied, we would argue that full autonomy is unlikely to ever be attained for medical AI systems. Humans will always be required for oversight of algorithmic interpretation of images and data in this sensitive realm.

References

[1] Shih-Yi Chien et al. “An Empirical Model of Cultural Factors on Trust in Automation”. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58 (Oct. 2014), pp. 859–863.

[2] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. “Explainable artificial intelligence: A survey”. In: *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE. 2018, pp. 0210–0215.

[3] L. H. Gilpin et al. “Explaining Explanations: An Overview of Interpretability of Machine Learning”. In: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. Oct. 2018, pp. 80–89.

[4] L. Clark Johnson et al. “Sampling bias and other methodological threats to the validity of health survey research”. In: *International Journal of Stress Management* 7.4 (2000), pp. 247–267.

[5] David Martens et al. “Performance of classification models from a user perspective”. In: *Decision Support Systems* 51.4 (2011), pp. 782–793.

[6] Tim Miller. “Explanation in Artificial Intelligence: Insights from the Social Sciences”. In: *CoRR* abs/1706.07269 (2017). arXiv: 1706.07269.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2016, pp. 1135–1144.

[8] Cynthia Rudin. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[9] Eric J Topol. “High-performance medicine: the convergence of human and artificial intelligence”. In: *Nature medicine* 25.1 (2019), pp. 44–56.

[10] M. B. Zafar et al. “Fairness Constraints: A Flexible Approach for Fair Classification”. In: *Journal of Machine Learning Research* 20.75 (2019), pp. 1–42.