Towards Patient-oriented Transparency

Short paper

Dina Babushkina[1][0000-0003-4899-8319]

[1] University of Helsinki, Helsinki, Finland

dina.babushkina@helsinki.fi

Abstract. The paper introduces a patient-oriented concept of transparency in application to medical AI systems. It is argued that the patient's perspective differs significantly from the user's perspective and that there is a need to re-think what transparency entails for a patient rather than for a professional user. It is suggested that there are two major factors that influence the concept of transparency if we take the patient into account: (a) the active patient paradigm and (2) the patient's trust in the medical sphere. It is further argued that, for an active patient, it is important to be able to make an informed decision about her health and voice criticism of a diagnosis/treatment, if she has sufficient ground for it. This poses one type of constraint on the concept of transparency. The other type of constraint derives from the fact that, in order to be trusted, medical expertise must be rooted in scientific methodology.

Keywords: ethics of Artificial Intelligence, transparency, black box, patient moral rights, medical expertise

1 Accuracy vs. transparency?

In his talk "AI and Trust: Explainability, Transparency" at Frankfurt Big Data Lab, Dragutin Petkovic posed a question to the audience, asking whether they would adopt a 99% accurate AI system which is essentially a black box, or a system that is less accurate (say in the range 85–95%) but explainable. At play here are a number of conflicting intuitions.

One of these intuitions feeds on—in the absence of a better word—hope. Hope that there must be this one magic solution to problems that we either cannot solve or solve imperfectly. A kind a methodological panacea, if you wish. This solution should come from a source that is better, smarter, more perfect than we are. The other intuition takes the form of a critical, nagging voice telling us that something which cannot be explained should not be trusted. It also says that there are no magic solutions, and the best way to address a problem is through the careful analysis of causal chains. Both intuitions are understandable and relatable.

But then, imagine going to a doctor. Say, you suffer from symptoms which are hard to diagnose. Your doctor uses an AI diagnostic system—let's call it the BDSE, the "Best Diagnostic System Ever"—and, based on the results, explains that you have X. You find this hard to believe. You find yourself doubting the diagnosis, because it contradicts your experience and your own assessment of your symptoms (to the best of your knowledge). You ask for an explanation, and you are told that you should not doubt the results, because the BDSE is known for its very high degree of accuracy. Would you change your mind, based on this reply? Or, even better, should you?

These are, however, two different questions. Whether you change your opinion is an idiosyncratic matter. You may do so for a number of reasons: because of a high level of trust in a professional medical opinion, because you cannot afford to pay for further examinations, or that you simply do not realize that disagreeing is an option. Whether you should be satisfied with the doctor's reasoning and, as a result, silence your doubts, should be considered on different grounds. It is rather a questions of whether there is a well-justified reason for you to do so, such that it would be right for anyone in your shoes to accept.

And how about the medical professional: should she accept the BDSE's result just because it comes from a system with 99% accuracy, even when this result is not fully in sync with what she can observe herself and with what her patient reports?

To make sure, the question being posed here is not whether "black-box" AI methods are useful, insightful, or whether we should use them. These questions are too broad. In many cases they are appropriate and we should use them. There is no objection, for example, when "black box systems" are used "as tools to inspire and guide human inquiry" and provided that, "as … any decision-making system, the black box … [is] used with knowledge, judgment, and responsibility" (Holm, 2019, 27). The question here, rather, concerns the ethical and rational constraints on their use. It is problematic to claim, for example, that black boxes *should* be used "when the cost of a wrong answer is low relative to the value of a correct answer" (Holm, 2019, 27). This is because the cost depends on the perspective. While, from the point of view of a medical institution, the cost of a low frequency mis-diagnosis might be low compared to the overall success of the AI-based diagnostic system, the cost is at a maximum for the person who is wrongly diagnosed and treated. The low-cost rational cannot be seen as a justifying reason[1] for the use of black-box methodology.

On the other hand, it is hard to deny that a black box may perform better than a human. For instance, "in reading standard field-of-view medical images, trained AI systems enhance the performance of human radiologists at detecting cancers. Although the cost of a wrong answer, whether a false negative or a false positive, may be high,

[1] More on justificatory (normative) vs explanatory (motivating) reasons see, e.g., Smith (2000, 95).

the black box offers the best solution that is currently available" (Holm, 2019, 27). This might be true, but only in a strictly controlled environment where the data, on which the system performs its predictions, correctly represents the training dataset. Black-box systems do not operate with the concept of causality; they rely on pattern recognition. For this reason, the environment in which the patterns are detected must be meticulously checked. For that reason, we cannot accept the claim that a black box *should* be used "when it produces the best result" as a justifying reason either (Holm, 2019, 27). My point is that there are good reasons never to uncritically accept unexplainable AI systems when this may entail risk to health and well-being of medical patients. In what follows I will touch upon two such reasons: the patient's active role in matters of her life and health and a patient's trust in medical expertise.

At the center of this discussion is the concept of transparency. By now it is apparent to many that transparency should be a requirement for AI systems (on transparency as a principle of Responsible Research and Innovation, see Dignum, 2019), especially in sensitive domains, when the output of such systems is used for decision making that affects well-being and poses a risk of harm. Whenever the need may arise to give an explanation for a decision, the possibility of explicating the role and contribution of an AI system must be provisioned. Transparency is an interesting and multi-dimensional topic. From the philosophical point of view, the first question concerns the kind of transparency needed and the nature of explanation itself (Coeckelbergh, 2020, 121). I want to bring in the perspective of the medical patient and see if this forces us to re-conceptualize transparency (on the importance of transparency in the medical sphere, see, e.g., Topol, 2019). *Does the patient's perspective pose specific constraints on the way AI systems should be transparent?* If yes, then how are we to understand transparency in order for it to respect the interests and well-being of a patient?

2 Constraints on transparency

My goal is to understand what the transparency requirement for AI systems in the medical sphere implies, when taking into account what matters for a patient.

Why is this issue important? It is important because most of the discussion on the transparency of AI methods is focused on the user and, as a result, on the knowledge that a user must be granted access to (see, e.g., Samek et al., 2017; Hagras, 2018; Petkovic et al., 2018). But when it comes to medical application, the patient is often not the user. Their relation to a specific AI system (be it a diagnostic system, therapy system, or an advisor system) is radically different than that of a user. Medical personnel, administrators, researchers, and, perhaps, developers count as users, when they employ a system as a tool in their work or research. This constitutes professional use, the goal of which is to enhance and aid performance. Patients are on the receiving end of this professional use; they are either those who benefit or are harmed due to the use of the tool by medical professionals. (I will leave out of this discussion for now the use of health gadgets, when patients may also be considered as users in such cases.)

One could ask, Why should we care about the patient's perspective? The patient's perspective matters for several reasons. I will mention two:

(1) Patients are those who are directly harmed in the case of an error. It is their life and health that is at stake, therefore it is only logical that their perspective is taken into account.

(2) Transparency is not a simple binary phenomenon. When it comes to complex systems such as medical AI, the question of transparency is inseparable from the questions as to what must be revealed, to whom it is revealed, what their background is, and for which purpose they need the information. Different bits of information are important for creating understanding in a user, such as the radiologist or pathologist (whose goal it is to make sure that there is a correct diagnosis and treatment); an AI researcher (whose goal is to ensure that the system produces the result it is intended to produce); and a person whose lungs the machine is analyzing and who needs to decide how to take care of his health and well-being.

Let me make a preliminary suggestion that two major factors influence the concept of transparency, if we take the patient into account:

(1) The contemporary paradigm of participatory medicine and the concept of an "active patient," and

(2) The patient's trust in medical expertise and the concept of good medical care (i.e., which practices must be considered acceptable as medical care).

3 Active patiency

The contemporary paradigm of participatory medicine presupposes that the patient and medical personnel work together to establish the best way to safeguard the patient's health.[2] This dramatically changes the roles of the patient, from being a passive receiver of care (whose role in decision making is more or less reduced to obeying or disobeying the medical authority) to an active agent, carrying the responsibility for her own health. An active patient is delegated a much richer decision-making role. Such changes in patient responsibilities entail changes in the patient rights and capacities. Most importantly:

[2] There are multiple sources on the topic, see, e.g., de Bronkart (2018), van der Eijk et al. (2013), Lejbkowicz et al. (2012), Hood and Auffray (2013, December 23).

(1) A right to make an informed decision in matters concerning her health and well-being. This presupposes that the patient must be able to make an informed judgment about her condition and state of health as well as the correctness of the diagnosis and effectiveness of treatment.

(2) A right to oversee the diagnostic and treatment processes and challenge the decision concerning her health, if she has sufficient reason to do so.

Neither of the patient's rights, however, can be realized, if the process of medical diagnosis cannot be explained. I do not mean that a person should not be given freedom to prefer a more accurate system to the one that is more transparent. It is similar to the freedom to choose an alternative treatment if that is more rational in their own eyes. What I mean is that one should not expect that every patient will make this choice. Non-transparent systems should not become default medical tools that are used without a patient's consent. In other words, acceptance of medical "black boxes" should not become a matter of the patient's silent agreement.

But what is it that matters from the patient's perspective? If transparency is understood in terms of code availability, then it is not the issue. While the availability of the code may be crucial for research and development purposes, when it comes to the patient's perspective, I do not think that making code known helps much. For one reason at least: it is almost impossible to read and interpret code without sufficient skills in programming.

There are other technical senses of transparency, which again target professionals, such as users and policy makers. None of them, I believe, is directly relevant to the patient's perspective. This is clear in the case of judicial transparency, which is understood as the requirement that "[a]ny involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority" (Petkovic, 2020). Another is failure transparency or the requirement that if an AI system caused harm, it should be possible to explain why (Petkovic, 2020). This is a rather narrow sense of transparency which demands an explanation post-factum, when the harm has already been done. What matters more for a patient is that there is a way to access the role of the system beforehand and prevent the harm. Such mistakes can become too costly for a patient.

Transparency as explainability, or the ability to provide an explanation of the procedure that an AI followed and the specific decision that has been produced as a result, comes closer to what should represent a patient's interests. But this is still problematic. To explain is not the same as being understood, and it is understanding that is crucial for the patient. One could, for instance, offer this explanation to a patient: the system used a pretrained deep neural network N to identify cancer and pneumonia patterns; its output was a set of values with corresponding percentages; the doctor based her diagnosis on the highest value output by the system. This explanation would constitute understanding (and contribute to a patient's ability to critically evaluate this result) only if the patient knows sufficiently well how such algorithms work, what their limitations are, and how to avoid misinterpreting the results. Most of us cannot do this. As a result, this explanation would not provide the right kind of knowledge for the patient to make an informed opinion. What matters for the patient is that she is given the right information that allows her to make an informed opinion such that she is in a position to challenge and object to an AI-based decision if she has sufficient reason to do so. And this bit is contained in the knowledge of the limitations of the AI method that influences the medical decision making in her case. Such limitations should ideally reflect how well the type of data that the model has been trained on represent the patient's case, since this is one of the key parameters of the model's accuracy (for a similar approach to data transparency see Yanisky-Ravid and Hallisey, 2018).

To sum up, I suggest that the patient should have access at least to the following:

- Information on which the AI method has been used for her diagnosis/treatment (so that she can refuse it or/and inform herself about the method from independent sources);
- Explanation of how this specific method works and how to correctly interpret its output;
- The limitations of this AI method.

4 Trust in medical expertise

But what about a frequent argument that since "we routinely accept human conclusions without fully understanding their origin" (Holm, 2019, 27), we should also accept AI unexplainable results? Pointing out that people casually accept unsupported conclusions is not much of an argument. Again, they might do so for multiple reasons: lack of time, trust in authority, optimism, laziness, or carelessness. But these are idiosyncratic reasons, which do not oblige everyone to make the same choice. Observing how people behave tells us nearly nothing about what we should or have a good reason to do, and about what counts as a well-justified practice for everyone. Indeed, the conclusion one uncritically accepts may be based on bad reasoning and therefore may be mistaken. Furthermore, there is a difference between "I accept X even though I do not fully know why X" and "I accept X even though no one is able to explain why X." While the first can still be justified in the case of well-established facts, the latter is esotericism, a mystical knowledge.

In this light, there is another important issue to consider: the patient's trust in medical expertise. This trust is crucial for the participatory medicine model to work. The foundation of this trust is the belief that medical institutions adhere to the principle of good medical and scientific practices. The patient has a good basis for trusting medical advice, iff[3] it is based on methods in line with established scientific standards. Among them is the requirement that a method or a tool is falsifiable, that is, there is a possibility of its refutation. It also means being aware and open about the imitations of the method, and the parameters that limit its applicability or challenge its use. These set additional scientific constraints on the concept of transparency.

A transparent AI system is such that:

- Its function is limited to that of a tool or method, aiding medical expert in her work;
- The medical expert is able to scientifically access the tool/method and its limitations and correctly interpret its results;
- The methods are such that they satisfy scientific constraints, and that implies that they are not obscure.

The ability for a researcher to evaluate an AI method and for a doctor to understand how it works as a medical tool (and to which extent its results have to be supplemented by expert opinion) are necessary for the patient to be provided with the required level of transparency. For a medical specialist employing an AI tool, the first question should not be about accuracy but about adherence to scientific procedures. This would allow her to provide adequate support in building the patient's trust. This, I must add, is not a problem of AI itself, but a much broader problem of the justifiability of a method.

I do not mean that there is no room for anything that is not fully explainable or scientifically sound. Sometimes people may be justified in turning to alternative methods, especially when the medical system fails them for one reason or another. Such methods also rely on methods that they are unable to scientifically account for. But the fact remains that these are esoteric practices that are not supported by the scientific paradigm and they should remain a matter of personal discretion. One is free to use them, provided that one is aware of their limitations and risks.

5 References

Coeckelbergh, M. (2020). AI ethics. Cambridge, MA: The MIT Press.

de Bronkart, D. (2018). The patient's voice in the emerging era of participatory medicine. The International Journal of Psychiatry in Medicine, 53(5–6), pp. 350-360.

Dignum, V. (2019). Responsible Artificial Intelligence. Cham: Springer.

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. Computer, Vol. 51(9), pp. 28-36.

[3] In standard logic notation, "iff" equals the biconditional "if and only if."

Holm, E. (2019). In Defence of the Back box. Science, Vol. 364(6435), pp. 26-27

Hood, L., Auffray, C. (2013, December 23). Participatory medicine: a driving force for revolutionizing healthcare. Genome Medicine. Retrieved from https://doi.org/10.1186/gm514.

Lejbkowicz, I., Caspi, O., Miller, A. (2012). Participatory medicine and patient empowerment towards personalized healthcare in multiple sclerosis. Expert Review of Neurotherapeutics, 12: 3, pp.343-352.

Petkovic, D. (2020, May 13). AI and Trust: Explainability, Transparency. [Online presentation]. Ethical Implications of AI and AI Tools Lab. http://www.bigdata.uni-frankfurt.de/data-challenge-ss-2020/

Petkovic, D., Altman, R., Wong, M., Vigil, A. (2018) Improving the Explainability of Random Forest classifier - User Centered Approach. Pac Symp Biocomput, 23, pp. 204-215. Retrieved from https://pubmed.ncbi.nlm.nih.gov/29218882/

Samek, W., Wiegand, T., Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. Retrieved from ttps://arxiv.org/abs/1708.08296

Smith, M. (2000). The Moral Problem. Oxford: Blackwell.

Topol, E.J. High-performance medicine: the convergence of human and artificial intelligence (2019). NatMed, 25, pp. 44-56. Retrieved from https://www.nature.com/articles/s41591-018-0300-7

van der Eijk, M, Nijhuis, F.A., Faber, M.J., Bloem, B.R. (2013). Moving from physician-centered care towards patient-centered care for Parkinson's disease patients. Parkinsonism Relat Disord. 19(11), pp. 923-927.

Yanisky-Ravid, S., Hallisey, S. (2018). ‘Equality and Privacy by Design’: Ensuring Artificial Intelligence (AI) Is Properly Trained & Fed: A New Model of AI Data Transparency & Certification As Safe Harbor Procedures. SSRN Electronic Journal. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3278490