Explainable AI in healthcare: why, what, and how technical potential could empower clinicians' capability.

Maura Bellio, Neil P. Oxtoby, Daniel C. Alexander, Ann Blandford

21th June 2019

The advent of big data and increasing potential of computing power has boosted opportunities for artificial intelligence applications, and particularly of machine learning (ML) models. We teach these systems how to make sense of data, when its amount goes far beyond human intelligibility. Their way to operate is inspired by our learning processes and brain development, where pieces of usable intelligence are built from previous experiences, repeatability and generalisation of real world phenomena [1]. Whilst these models might help us making sense of big data, then **why** do we need explanations? We need them due to the opaque nature of ML, to support fundamental requirements, such as transparency, fidelity, and trust [3]. *Explanations* are defined by Lombrozo as "the currency in which we exchange beliefs" [2]. From a ML perspective, that currency is represented by translating AI procedures into "understandable terms to a human" [4].

A domain area where the demand for explainability is particularly critical is healthcare. This field takes advantage of AI for a number of tasks, such as decision-making, predictions, risk management, and policy [5]. My research focuses on how to bring ML models for *prediction* of disease progression to clinical practice. Predictions are a delicate theme, as they can have far-reaching implications. Imagine how automatic predictions might guide treatments' decisions in the future. To date, clinicians normally make predictions based on recurrent (mostly qualitative) observations, expertise, or even general impression. Given new potentials in healthcare, clinicians have to make their way through an overwhelming number of variables. Thus, we handcraft systems to overperform human intelligence and be able to see trends and patterns the human eye would not capture at a glance. We also aim for higher performance, but the trade-off is that more interpretable models, such as regression models or decision trees, normally offer lower performance [6]. Therefore, how can we claim that we need explanations more than we need performance? I believe the problem cannot be reduced to this. We are not expecting humans to grasp that sophisticated calculations for which they built computational systems as an aid. In fact, explanations should be tailored on **what** matters to the person, and not on a deep understanding of the model [7]. One way to do that could be by interpreting the outcome, and understanding when the system works or not. For example, I care that my car moves, stops, and steers as I need, in a safe and efficient way, but not really about the engine's functioning details. So, what makes me trust my car? It is not only confidence in the manufacturing process, but also my ability to promptly detect when something is not working. Similarly, ML interpretability is not necessarily about a deep understanding of the algorithms, rather than being able to audit feedback on the intended performance and trust in the construction and validation process.

This leads to **how** explanations can promote trustfulness, for which three points can be made:

- Account for generalisation errors;
- Be role-dependent;
- Provide a user interface.

A good balance between performance and transparency might lie in how we learn to identify *deviations* produced by the AI system, also known as "generalisation errors" [3]. Think of IBM Watson [8], easily beating the most skilled chess player, whilst completely failing when recommending cancer treatments. Defining cases for when a system does not perform, allows the user to decide to what extent it is fine to trust its suggestions. In other words, we want to be able to define ML *uncertainty*. Some research is looking into training models to provide their uncertainty quantification, and the major underlying causes for it [9]. Explanations should also be *role-dependent*, reflecting the intended use and users of AI in the specific context. In role-based scenarios, the level of predictions' explainability required by a physician will inevitably be different from that of a staffing planner, despite both being part of the same domain [3]. This suggests that ML outcome should be designed based on users' needs and workflow [3]. A good strategy is that of using domain-specialised terminology, to make the interaction intuitive, familiar, thus fostering interpretability. This bridge between clinical practice and AI can be further strengthened by efficient *user-interfaces* [1], generally in the form of graphical platforms mediating this interaction. Interfaces offer new strategies of translating AI to healthcare practitioners, providing users with a tool that encourages the exploration of debugging [1], or more generally an understanding of a system's accuracy in actionable predictions. With most AI models in healthcare being limited to offer predictions only, what represents a future towards explainability is to pair AI outcomes with medical reasoning on possible actions, as influenced by the interpretable outcome [3].

[1] 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 preprint arXiv:1712.09923*.

[2] Lombrozo, T. (2006). The structure and function of explanations. *Trends in cognitive sciences*, *10*(10), 464-470.

[3] Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable Machine Learning in Healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 559-560). ACM.

[4] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[5] Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719.

[6] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., ... & Wood, A. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint*.

[7] Cassie Kozyrkov. (2018). Explainable AI won't deliver. Here's why. Retrieved from: https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be

[8] Ross, C., & Swetlitz, I. (2018). IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Stat News https://www. statnews. com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments*.

[9] Tanno, R., Worrall, D. E., Ghosh, A., Kaden, E., ... & Alexander, D. C. (2017). Bayesian image quality transfer with CNNs: exploring uncertainty in dMRI super-resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 611-619). Springer, Cham.

word count:
997 (including references)