Explainable AI is indispensable in areas where liability is an issue

Nelson Brochado

June 17, 2019

Certain animals and, in particular, humans have always been curious about the mysteries of the world. We have always shown interest in exploring the unknown, so that it becomes known. The necessity of discovery is likely inherent to our nature and it is possibly related to our limited time. Throughout the years, we have developed ways of communicating with each other and other animals. In particular, we have developed ways of saving and transferring information and knowledge. We have also developed ways of automating certain tasks, notably artificial intelligence, which is one of the most promising fields, given that it has the potential to enhance the quality of our lives. However, AI has also a few limitations and can also be dangerous. So, there is the need, more than ever, to attempt to solve these limitations and avoid the dangers.

Explainable (or interpretable) artificial intelligence (XAI) can refer to techniques that are used to explain (to humans) or interpret the inner workings and outcomes (e.g. predictions) of *black box* artificial intelligence (AI) or machine learning (ML) models. XAI can also refer to *white box* AI or ML models whose inner workings and outcomes (or outputs) can be more easily (with respect to black box ones) understood and trusted by humans [1] [2].

XAI is related, in several ways, to strong AI (or artificial general intelligence), causation, statistics (e.g. visualisation techniques), credit assignment problem, human-computer interaction, cognitive science, neuroscience, psychology, linguistics, consciousness, meta-learning, legislation and ethics (e.g. free will).

[1] The definition of XAI is not yet standardised. For example, [1] distinguishes between interpretable and explainable (or explanatory) AI. In this article, I am using explanatory (or explainable) and interpretable AI interchangeably.

[2] Funnily enough, the expressions *black box* and *white box* are, in a certain way, discriminatory. However, humans tend to categorise and give meanings to categories, which can be criticised (for several reasons) by other individuals who have not created those categorisations or simply do not agree with them.

Several XAI techniques have been proposed in recent years. For example, *layer-wise relevance propagation*, which allows to visualize (as heat maps) the contributions of single pixels of an image (which, in this case, is the input to the model) to predictions for multi-layered neural networks (and kernel-based classifiers over bag of words features) [2, 3], and *LIME*, which explains the predictions of any classifier in an *interpretable* and *faithful* manner by learning an interpretable model locally around the prediction [4]. These XAI techniques can be considered (partially) *model-agnostic*, given that they can be applied to more than one model. There are other possible categories and categorisations of XAI algorithms and models. For example, there are model-specific or example-based XAI techniques [5].

The typical examples of black box AI or ML models are artificial neural networks (ANNs), which, in the last decade, have achieved state of the art performance in several tasks, such as machine translation [6], but that are still not well understood or, at least, not as much as sometimes is required. An example of a white box model is the decision tree, whose inner workings and predictions have a relatively clear interpretation or explanation (to humans).

ANNs have been shown to be strongly biased towards certain outcomes [7, 8] or easily fooled [9]. Several works studied the limitations and robustness of ANNs [10, 11, 12, 13, 14, 15]. These works have shown that ANNs are sensitive to small perturbations of the inputs (which are e.g. images or text), that is, the output of these ANNs is highly affected by small perturbations or changes of the inputs[3]. Furthermore, several works have also shown that ANNs can be easily attacked or hacked [16, 17].

There are a few questions regarding these limitations and issues of ANNs that need to be answered. Are these problems due to wrong inductive biases of the models [18] or due to the use of non-representative data of the population (or both)? Can an ANN perform some operation that is not understandable or conceivable to humans, so certain ANNs will always be black boxes? If yes, how should we deal with these issues? What is the relation between current ML approaches and causation (or causal inference)? Is causal inference strictly required to be implemented or integrated into these systems so that to avoid these issues? Is there an a trade-off between performance and interpretation of the inner workings of a model?

Nevertheless, in theory, ANNs (with at least a single hidden layer containing a finite number of neurons) can approximate a wide variety of *interesting* functions[4], so they are often denoted by *universal function approximators* [20, 21, 22, 23, 24]. Their practical performance is thus also a consequence of their theoretical *powerfulness*, but it is also due to the advances, in the last decades, in computer hardware. There are also other universal function approximators, for example, support vector machines [25], but, in recent years, they have not received as much attention as ANNs, even though they are still a relatively valid ML method.

[3]Perturbed inputs are called often called *adversarial inputs*.

[4]More specifically, these interesting functions are continuous on compact subsets of $\mathbb{R}^n$ [19].

In areas that involve human (but not only) lives, ethics or decisions that can severely impact the life of people (and other living beings), like in healthcare, autonomous vehicles, military missions, banking, insurance, legislation, jurisdiction, climate change, safety or security, the adoption of ANNs can lead to undesirable and unfixable consequences, because of their limitations and weaknesses mentioned above, so some people are legitimately hesitant to adopt these black box ML techniques, even after considering their potential benefits and their recent successes.

For example, in the context of healthcare, an explainable AI model that is in charge of prescribing a dose of a certain drug could provide an detailed explanation of such decision to the doctors, so that the doctors could first discuss it and eventually approve it or not. If such an explanation is not available, the doctors can still hypothesize the reasons behind such prescription, but, under the hood, the prescription by the AI might be due to completely different reasons, which the doctors might not have envisaged. Consequently, the doctors could be misleadingly persuaded by the unexplainable decision of the AI to take a certain possibly fatal action. In the case the doctors do not have the time to discuss the AI prescription, they can either accept or not the prescription. Who will be responsible for the possible undesirable or fatal outcome?

Can an AI be responsible for such fatal or undesirable outcome? Can an AI have free will and rights (in a similar way humans do [26])? There are futurists who claim that artificial general intelligence and even super-intelligence is possible and that we should be concerned with the possible associated dangers [27]. However, at the moment and in the near future, the answer to these questions is unlikely to be affirmative, given that most people are not fond of the AI field to the point of conceding it free will and rights, because AI today is still quite inflexible and not general. Furthermore, AI does not possess the characteristics of a living being, so, considering that many humans do not even concede rights to other living beings or even humans, it is really unlikely that they will concede them to inanimate entities (like an AI).

The specific answer to all these questions depends on the actual legislation, which humans must establish (in a process that possibly involves an AI agent). Given the mentioned weaknesses of ANNs and the possible undesirable consequences of the outcomes of their use, organizations, like the European Union, start to be concerned with these issues. More specifically, the European Union's recent General Data Protection Regulation [28] introduces new laws that attempt to protect European citizens against the possible incompetence of AI or ML models. In particular, the "right to explanation" of algorithmic decisions attempts to protect the European citizens against these issues that can arise from the use of unexplainable AI models [29] (and other related issues). Nonetheless, legal scholars have already criticised this right and stated that it is unclear and unlikely to present a complete remedy to algorithmic harms [30]. More specifically, an appropriate explanation depends on the context or situation and can be subjec-

tive, so the promulgation of such related laws is not an easy task, considering that the AI field is still evolving and its consequences, dangers or threats are not well understood yet. Regardless of the defects of this and other laws, the main point is that certain important organisations start to care more about the rights of humans with respect to the invasion of technology and, in particular, AI, and the possible undesirable consequences of its adoption.

To conclude, the concept of explainable AI becomes always more relevant due to the recent successes of AI and, more specifically, deep learning techniques, which arouses interest in adopting these techniques to automate certain tasks (that involve ethics and lives), which are currently mainly performed by humans. For this reason and other reasons, several important organisations, like the European Union, have started to promulgate laws that attempt to protect citizens against the invasion of technology and the related incompetence of humans. Finally, the creation of a *satisfactory* XAI might be an AI-complete problem (that is, a true XAI might be equivalent to an artificial general or human-level intelligence [31]), so, in other words, this is a hard problem.

References

[1] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning," *arXiv e-prints*, p. arXiv:1806.00069, May 2018.

[2] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, pp. 1–46, 07 2015.

[3] A. Binder, G. Montavon, S. Bach, K.-R. Müller, and W. Samek, "Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers," *arXiv e-prints*, p. arXiv:1604.00825, Apr 2016.

[4] M. Tulio Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *arXiv e-prints*, p. arXiv:1602.04938, Feb 2016.

[5] C. Molnar, *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv e-prints*, p. arXiv:1706.03762, Jun 2017.

[7] S. M. Julia Angwin, Jeff Larson and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.," *ProPub-*

lica, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[8] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, pp. 183–186, Apr 2017.

[9] A. Nguyen, J. Yosinski, and J. Clune, "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images," *arXiv e-prints*, p. arXiv:1412.1897, Dec 2014.

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv e-prints*, p. arXiv:1312.6199, Dec 2013.

[11] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," *arXiv e-prints*, p. arXiv:1610.08401, Oct 2016.

[12] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. Berkay Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," *arXiv e-prints*, p. arXiv:1511.07528, Nov 2015.

[13] R. Jia and P. Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems," *arXiv e-prints*, p. arXiv:1707.07328, Jul 2017.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv e-prints*, p. arXiv:1412.6572, Dec 2014.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *arXiv e-prints*, p. arXiv:1706.06083, Jun 2017.

[16] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25nd Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-221, 2018*, The Internet Society, 2018.

[17] J. Clements and Y. Lao, "Hardware Trojan Attacks on Neural Networks," *arXiv e-prints*, p. arXiv:1806.05768, Jun 2018.

[18] J. Baxter, "A Model of Inductive Bias Learning," *arXiv e-prints*, p. arXiv:1106.0245, Jun 2011.

[19] Wikipedia contributors, "Universal approximation theorem — Wikipedia, the free encyclopedia," 2019. [Online; accessed 16-June-2019].

[20] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, pp. 303–314, Dec 1989.

[21] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251 – 257, 1991.

[22] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 6231–6239, Curran Associates, Inc., 2017.

[23] B. Hanin, "Universal Function Approximation by Deep Neural Nets with Bounded Width and ReLU Activations," *arXiv e-prints*, p. arXiv:1708.02691, Aug 2017.

[24] B. C. Csáji, "Approximation with artificial neural networks," *Faculty of Sciences, Etvs Lornd University, Hungary*, vol. 24, p. 48, 2001.

[25] B. Hammer and K. Gersmann, "A note on the universal approximation capability of support vector machines," *Neural Processing Letters*, vol. 17, 10 2002.

[26] United Nations, "Universal declaration of human rights." `https://www.un.org/en/universal-declaration-human-rights/index.html`, Dec. 1948. Accessed 17 June 2019.

[27] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. New York, NY, USA: Oxford University Press, Inc., 1st ed., 2014.

[28] E. Commission, "2018 reform of eu data protection rules." `https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en`, 2018. [Online; accessed 17-June-2019].

[29] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *arXiv e-prints*, p. arXiv:1606.08813, Jun 2016.

[30] L. Edwards and M. Veale, "Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for."

[31] D. Shahaf and E. Amir, "Towards a theory of ai completeness," in *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2007.