XAI: Digital Ethics

Adriano Koshiyama, Emre Kazim, and Zeynep Engin

Department of Computer Science, University College London

E-mail: akoshiyama@cs.ucl.ac.uk ekazim@cs.ucl.ac.uk

Submitted as position paper to: Human-centred explainable Artificial Intelligence (HeXAI) Workshop (July 5th 2019, University College London).

Abstract

In this positon piece we present our current thinking concerning Explainable Artificial Intelligence (XAI). We state our interest and provide the broadstrokes of our approach and questions that we are considering. We hope this will function as a point of departure for discussion with other invitees.

Keywords: Explainable Artificial Intelligence, XAI, Digital Ethics, Transparency, Accountability, Explainablity, Fairness, Robustness

1. Introduction & Background – Digital Ethics & AI

We are interested in:

i. social and political concerns around new data science technologies with particular focus on disruptions occurring around their deployment potentials in the public sector.
ii. technology in relation to 'policing' safety, fairness, trust, legality, interpretability and performance of algorithmic decision-making and support systems in generic terms, introducing the context to AI explainability in financial markets.

Our understanding of AI Transparency is reducible to three parts:

i. explainability,
ii. interpretability,
iii. accountability and traceability

Currently there are no consensual definitions of these (from a computer science perspective). Our approach involves the following topics/criteria that can assess an AI / Machine Learning system level of transparency:

- AI Explainability: (i) directly interpretable machine learning vs post hoc interpretation; (ii) different users' needs from explanations; and (iii) the current quest for a satisfactory quantitative definition of interpretability; and (iv) recent regulations in the European Union requiring 'meaningful' explanations.
- AI Traceability: basically, this topic deals with ensuring that we can track and maintain the provenance of datasets, metadata, models along with their hyperparameters, and test results. Users, those potentially affected, and third parties, such as regulators, must be able to audit the systems underlying the services. Appropriate parties may need the ability to reproduce past outputs and track outcomes. A good deal of literature is available on Data Provenance, but AI Provenance is still a research topic in its early stages.
- AI Accountability: the legal and ethical discussion about AI Accountability will be of particular interest of the regulators. What are the concrete means in which an individual can hold an algorithm to account? For instance, a company having legal personality can, for example, be charged with corporate manslaughter, which is a criminal offence in law being an act of homicide committed by a company or organisation. Other legal aspects, such as of Agency, where a relationship is created where a principal gives legal authority to an agent to act on the principal's behalf when dealing with a third party, are also a concern.

2. AI / Machine Learning Explainability: concepts, techniques and tools

We have a strong interest in focusing on the main concepts, techniques and tools that enable opening the black-box of an AI / Machine Learning method. More specific on describing and categorising the techniques and tools available into a four-quadrant chart: Global and Local (horizontal axis) and Model-Specific or Agnostic (vertical axis) (Figure 1).

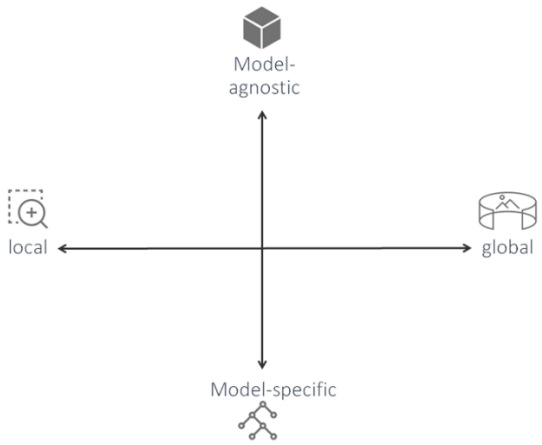

Figure 1. Types of explanations.

This grouping allows to quickly localize and understand the difference between the plethora of methods available, as well as provide a roadmap for their description. After this categorisation, we can drill-down in the specifics of each method, such as:

i. types of explanation they provide – feature statistic or visualisation, model-internals, rule-based models and counterfactual explanation;
ii. their acceptance and maturity in the AI / Machine Learning community – a broad bibliometric review in main conferences and journals);

iii. easiness to implement/reuse implementations – an in-depth analysis of the use cases, open-source platforms, tools, languages, etc.; and

iv. to deal with time-varying scenarios – updating explanations in real-time, providing insights of what is driving changes in the decision-making process, and so on.

Also, for each topic, try to identify a potential case studies like: facial recognition, speech analysis, credit scoring, etc.

3. Moral/Philosophical concerns

We have questions concerning the extent to which explainability, traceability and accountability can in themselves be considered definable:

Explainability

i. What counts as an explanation? Identifying 'cause' / 'correlation' is problem of all social sciences: when identifying one or the other (or both), within a system, substantive 'social scientific' (statistical analysis), is required. These are debated and delve into significant points of methodological contention.

ii. Opacity thesis: we are unaware of ourselves. The object self and the subject self – AI actual – AI explanation of itself. There is a philosophical view that we can never access (opaque) our true self (motive, ratiocination etc.). Indeed, there are psychological/psychoanalytical empirical studies that support this. As such, a person's explanation can only concern the 'apparent' – by analogy, it may be that explanation in AI is only that which is apparent (it may be that objectively valid explainability is beyond us – and this may be fine). This allays with explainability – if we 'see' a particular bias (such as gender or racial) we may 'explain' results with respect to these factors, when in fact something else is taking place.

iii. How might we evaluate our evaluation?

Traceability

i. What are we tracing? When a system changes over time, we need to think about
   1. Data set expansion (Data set precision, Data set trimming i.e. removal of redundancies, etc.)
   2. Parameter change (i. changes in parameters, and ii. evolution of parameters).

ii. Tracing problems of obfuscation:

A1 (time 1) = Data Set (D1) + Parameter/model (P1) + ….

A2 (time 2) = (D1 + D2) + (P1 or P2) + … [- think of all the combinations of these]

Ex.1: If the algorithm at t2 is different only as a result of an introduction or edit of a parameter then D will = D1 (vice versa for P), however if algorithm 'learns' then D and P are unchanged and yet moved form A1 to A2 . What are we tracing, date, parameter, evolution?

Ex2.: if D changes, but we have introduced this amended data-set at point A2 – are we tracing from A1 or is A2 the new A1? What is our start-point? Does a new data set, or amended data set, count as a new starting point? Similarly, P1, and P2.

i.e. what counts as a transition from A1 to A2?

Ex3.: the algorithm evolves (learns), thus D and P remain unchanged, but A does change. We will have

A2(time2) = A1 (time2). i.e. are we tracing with respect to time only? And by time, do we trace the changes in the conclusions of the A? etc.. In cases of time stamping, how do we decide what 'snapshot' to take.

Accountability

i. Moral, legal and ethical questions regarding cases of unintended consequences? At what point can we hold some one, some company accountable for something beyond a reasonable level of risk analysis? More specific, how can the engineer be held responsible for this? Are these translatable to engineering problems?

References

[1] Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. "A survey of methods for explaining black box models." ACM computing surveys (CSUR) 51, no. 5 (2018): 93.

[2] Russell, Chris. "Efficient Search for Diverse Coherent Explanations." arXiv preprint arXiv:1901.04909 (2019).

[3] Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721-1730. ACM, 2015.

[4] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

[5] Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. "Explaining Explanations: An Overview of Interpretability of Machine Learning." In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80-89. IEEE, 2018.

[6] Mittelstadt, Brent, Chris Russell, and Sandra Wachter. "Explaining explanations in AI." arXiv preprint arXiv:1811.01439 (2018).

[7] Doshi-Velez, Finale, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. "Accountability of AI under the law: The role of explanation." arXiv preprint arXiv:1711.01134 (2017).

[8] Rudin, Cynthia. "Please stop explaining black box models for high stakes decisions." arXiv preprint arXiv:1811.10154 (2018).

[9] Lipton, Zachary C. "The mythos of model interpretability." arXiv preprint arXiv:1606.03490 (2016).

[10] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).