31.05.2019

HExAI Workshop Position Paper

# Reflections on Interactive Multimodal Approaches to Making Algorithms Explainable for Diverse User Groups

by *Cagatay Turkay*

When thinking about eXplainable AI, I like thinking back the great papers by Galit Shmueli [1] and Leo Breiman [2]. Shumeli talks about two different goals in computational modelling: explanatory modelling where the goal is to derive and test causal reasoning around an hypothesis, and predictive modelling where a computational model is applied to data for the purposes of predicting new/future observations. Breiman adopts a general yet formal definition and considers "nature" as a black box model where a set of (independent) variables go in and a response comes out on the other side. He then takes a similar position and talks about two goals when analysing data that tries to approximate nature: prediction and information extraction (information that helps understand how nature is associating the response variables to the input variables). I see these two different roles of AI as a good starting point to think about what we expect from data intensive computational models (AI?) and I think explainability has different significance for both of these roles. When the role of AI is prediction – most common use of AI/ML at the moment, e.g., mortgage loan decisions – explainability is mostly related to understanding the decisions (outcomes) made by the algorithms. XAI in this setting is primarily about the explainability of "machine decision-making" and commonly facilitated through investigating the factors (inputs) that have an impact on an algorithmic outcome. Unless it has a particular significance in understanding the end results, there is little interest in understanding the inner-workings of an algorithm in these cases. In other words, leaving the model as a black box is not necessarily a concern in this scenario. When the second (and maybe less common role especially outside scientific circles) of AI/ML systems of helping understand phenomena better is considered, expectation on explainability is more comprehensive. The explainee (as used by Miller [3]) is not only interested in understanding why particular decisions are made by an algorithm but has also a genuine interest in understanding the "sense-making logic" of the algorithm, i.e., the inner workings and logical constructs. The explainability criteria in this case thus has to cater both of these needs, i.e., not only make the relation between the inputs and outputs clear, but also communicate how these elements interact with each other and with the logical constructs of the algorithm that is processing the input data. Having made these distinctions, it is fair to say that these notions of explainability are overlapping and any approach to facilitate explainability in any of these settings can be beneficial for the other.

My research is in the area of data visualisation and interactive data analysis. I am interested in designing approaches where analysts interact with algorithms for the purposes of understanding the underlying phenomena in the data better or to build better data-intensive computational models. Although explainability has not been the main driver of some of the early research we did, we have seen that making algorithms interactive has facilitated better engagement of analysts and empowered them in utilising algorithms more effectively [4]. Through interactive interfaces, we always aim to elicit experts' tacit knowledge regarding a domain and pass that knowledge to algorithms to create models that are "informed" by their users' knowledge. This agency we provide to analysts not only leads to better models but to models that the experts feel ownership towards and trust in ways that they can defend to their peers. Looking back at this research from an explainability perspective, trying to make algorithms responsive and interactive has inevitability served into making these algorithms explainable as well.

More recently, we started investigating the use of multimodal representations [5] – a mix of data visualisation and natural language – for the purposes of explaining algorithmic results. We are

investigating how we can combine visual and verbal descriptions of algorithmic results. We try to develop a design space that enables us to reflect on this multimodal landscape and investigate how explanations can be generated and how explanations can be presented (Figure-1). In a follow up and ongoing work, we are investigating how we can generate "algorithmic stories" for the purposes of explainability through the use of narrative techniques and narrative templates that combine visual and textual representations [6] (Figure-2).

There are several questions I would like to think about going forward and interested in studying further:

- Investigate the relation between explainability and trust. To what extent explainability leads to trust in AI/ML? Do we feel more empowered by knowing more about the decisions, or will we trust them less as people discover how "naive" the algorithms are despite all their theoretical elegance?
- Investigating the role of enhanced communication between algorithms and humans. How can we design algorithms that learn from us and develop with us, as well as "we" learning from them? And what will that collaborative learning lead to in terms of our perception of algorithms?
- How can explanations be suited to different audiences and how information can be layered in ways that are adaptive to the needs of the explainees?
- What are some communication mediums (text, visuals, sound …) that we can leverage to develop more explainable models?
- How can we tackle the myths around AI and ML? How can we foster a healthy culture in society that is interested in understanding the strengths and the limitations of algorithms, embrace their inherent uncertainties, and eventually build realistic, critical, but still useful relations between humans and algorithms?

[1] Shmueli, G., 2010. To Explain or to Predict?. *Statistical Science*, *25*(3), pp.289-310.
[2] Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), pp.199-231.
[3] Miller, T., 2018. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.
[4] Turkay, C., Slingsby, A., Lahtinen, K., Butt, S. and Dykes, J., 2017. Supporting theoretically-grounded model building in the social sciences through interactive visualisation. *Neurocomputing*, *268*, pp.153-163.
[5] Sevastjanova, R., Becker, F., Ell, B., Turkay, C., Henkin, R., Butt, M., Keim, D. and Mennatallah, E.A., 2018. Going beyond Visualization. Verbalization as Complementary Medium to Explain Machine Learning Models.
[6] Liem J., Henkin R., Wood J., Turkay C., 2019, A Descriptive Framework of Stories of Algorithms, EuroVis 2019. (https://algostories.github.io/)

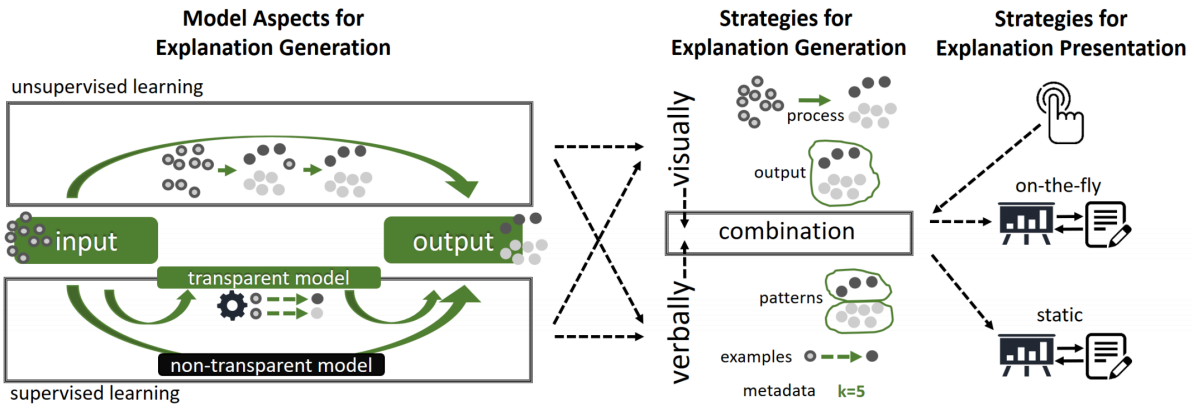


*Figure 1 A proposal for a design space for combining visualisation and verbalisation for model explainability [5]*

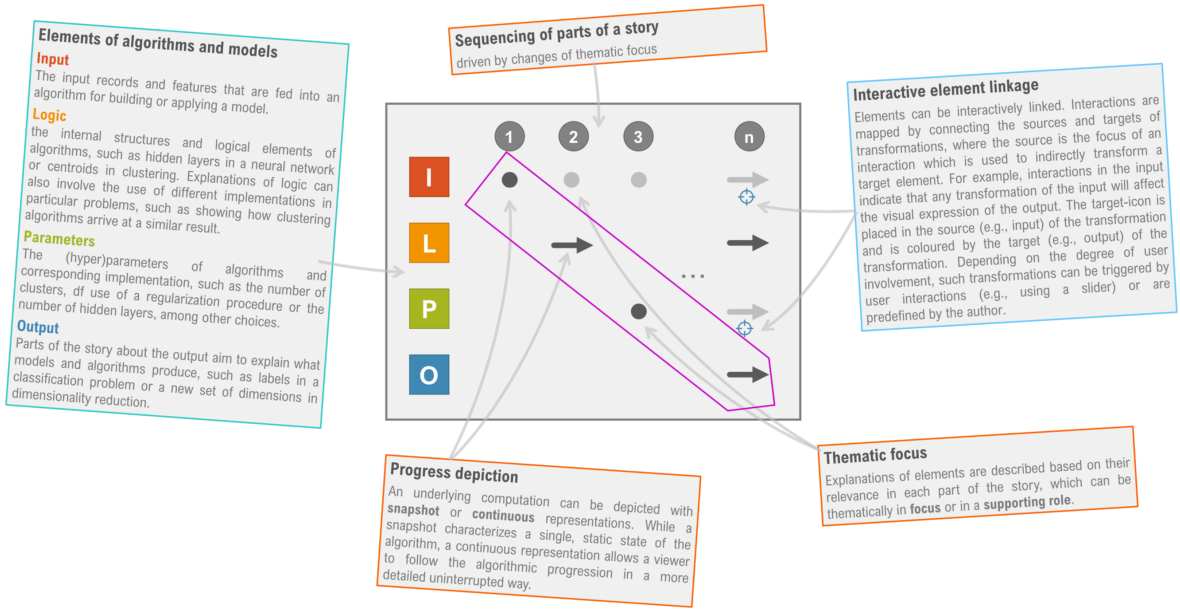


*Figure 2 A Descriptive Framework for Stories of Algorithms [6]*