25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Mind the gap: Making explainable AI better understood in contexts

Carolyn Huston

Data61, Commonwealth Scientific and Industrial Research Organisation
Email: Carolyn.Huston@csiro.au

Abstract: There is a gap in explainable machine learning (XML) and the ability of communities and contexts of application to understand XML and other explainable AI (XAI) methods. This is contrary to ethical and responsible innovation goals which emphasize explainability, contestability, transparency, interpretability and similar as necessary to responsible AI development (Arrieta et al. 2020; Department of Industry Science and Resources, 2021). Arrieta et all. 2020 provide a sensible backdrop of various purposes to explainability, as well as taxonomies to understand types of explanation models are capable of, for example being interpretable by design (as in a regression model) as opposed to interpretable by external techniques such as feature relevance computations. Despite this, many so called explainable methods are only meaningfully explainable to those with substantial AI expertise already; creating an ever-growing understanding gap in society where there is a fast-expanding AI audience as new AI tools are released permeate everyday activities, but where users are not AI experts.

Design research and thinking is an existing literature, research area, and body of practice that has expertise in understanding and deploying technologies in contexts (Norman, 2013, Davis, 2020). In the language of design, much of the current XAI literature focuses on complex *affordances*, where affordances represent what a model can do or be used to understand – XAI models *can* be understood or used to develop explanation or inference. Much explainable AI focuses on *affordances* in terms of the useability and understandability within the AI and ML communities. While some of the models and methods for explaining them are very powerful, there is often a breakdown in appropriately communicating the utility of such models in other contexts where model users are rarely AI and ML specialists.

Considering the many applied contexts where AI models are used, XAI literature does not consider a related component of design thinking, the development of *signifiers*. Signifiers indicate appropriate use of a technology in a context. They should indicate where and when and how an activity using the AI technology should take place for it to work properly, and to highlight when use of a model is risky or requiring unmet assumptions. Currently, XAI tends toward wordy explanations or require the ability to decipher algorithmic or mathematical symbology to communicate assumptions. These are legitimate signifiers but XAI often seems to be missing, or not applying thoughtfully, simpler signifiers. Think of the contrast in digestibility between a long-winded user-manual for a car vs. the relative simplicity of most controls within it. Building confidence and understanding of the application of the model requires thoughtful communication about limitations as well as understanding of cognitive load of those using it.

Task analysis (and subsequent human-centered design of XAI models and reporting) could be useful. Knowing what and how tasks disrupted by AI systems are performed *in-situ* could allow more appropriately explained models and signifers. Successful XAI requires design process and thinking in parallel with the development of models themselves.

REFERENCES

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, Francisco Herrera, 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82-115.

Davis, J. 2020. How artifacts afford: The power and politics of everyday things. MIT Press.

Department of Industry, Science and Resources, 2019 Australia's artificial intelligence ethics framework.

Norman, D. A., 2013. The design of everyday things. MIT Press.

Keywords: *Explainable machine learning, responsible innovation, affordance, signifier*