Thoughts on ITL's Draft Paper on ML Explanation

Peter Denno, NIST

2020/10/14

Abstract

Hi, I'm in EL, Systems Integration Division, where I'm finishing up a PhD on joint cognitive work (JCW). (I've been at NIST for many years, though.) JCW is about AI and people working together on cognitive tasks (as you might if you use tax preparation software to help do your income taxes). Explanation is an important part of JCW. I don't use neural nets for JCW but Bayesian nets, which, of course, more directly explain. I have studied some of the philosophy of science of explanation (and that of causality, which can be related). My comments below are likely to reflect that bias, and also a systems engineering perspective. I hope they are helpful. I'm around, of course, if you'd like to meet virtually.

1 Types of Explanation

I would suggest that what you are describing in *Section 3, Types of Explanation* as types of explanation are better described as purposes of explanation (in the sociotechnical context in which the system operates). Since the same explanation could serve more than one of these purposes, I do not see how these could serve as useful taxonomic types.

The philosophy of science literature on explanation has, over many years, developed several theories of explanation. Each concerns a different way of explaining (i.e. a type of explanation). Because these might be useful to explaining ML decision making, I describe some of the major ones below [1].

Deductive-nomological (DN) is a sound predicate logic argument based on premises which include necessarily at least one natural law. (See for example, Hempel [2] from the 1960s).

Statistical relevance is an argument based on statistical relevance of attributes, viz. $P(B|A \wedge C) \neq P(B|A)$ and homogeneous partitions of the attribute A into exhaustive and mutually exclusive subclasses [3].

Unification values the ability of an argument to unify theories.

Pragmatic is an account that simply answers a why-question.

Causal-mechanical is an account based on singular causal description grounded in physics.

Mechanistic "Mechanistic explanations appeal to the parts, operations, and organizations of mechanisms to explain the phenomena for which they are responsible [4]." In particular, Machamer, Darden, Craver (MDC) [5] is an account based on mechanism. Machamer et al. states "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions." [5] The MDC account additionally uses constructs referred to as mechanism sketches and mechanism schemata. The use of graphical constructs is typical of these more recent accounts of explanation.

Causal Intervention is an account that uses causal networks, counterfactuals and interventions (studies of effects holding all but one putative cause unchanged). It has been implemented in Bayesian networks [6].

Regarding these, causal intervention, emphasizes counterfactuals and methods such as D-separation; it is explanation geared for scientific inquiry. Conversely, mechanistic explanation would have appeal where ML is used to diagnose system failures, among other uses. You make the excellent point that explanations aren't one-size-fits-all. I'd add that it would be worthwhile to look into these different kinds of explanation and pair them with the needs of various stakeholders. This would be part of a recommended best practice in engineering AI systems.

2 Other Bodies of Knowledge

From what I gather from the ML literature (I'm not as engaged in it as you all), in practice to date, explanation is mostly about identifying features salient to local explanation. But people benefit from seeing how the features fit together and sometimes on seeing why other possible decisions and their rationale were discounted. They benefit from possessing inter-explanatory relationships, and they benefit from explanations grounded in the real world[1], rather than the model. I suppose that the way we move be-

[1] "real world" being either of physical phenomena (e.g. for diagnosis) or fiat business criteria (e.g. for credit worthiness).

yond the current situation is to (1) carefully assess what constitutes a good explanation in various settings, and (2) match explanation to stakeholder needs. A few paths forward come to mind; these are discussed below.

2.1 Philosophy of Science

Appeal to what the philosophy of science has to say about explanation. As I think is mostly evident from the definitions of different types of explanation above, "scientific explanation" need not use all the tools of science. I take the usage of "science" with a grain of salt. A theory is a hypothesis about the relationship between a model and reality [7]; the model can be a mental model.

Of course, there is a lot of literature here that won't be particularly relevant, but cherry-picking the best would be effective. For example, Khalifa's recent book [8] develops a theory about the properties of good explanation and how to assess whether one explanation is better than another.

2.2 Causality

Dig deeper into causality, an area that has developed markedly in the past few decades. Work on Bayesian techniques like you cite (e.g. Lethham et al. [9], Bayesian Rule Lists) I read as referencing causes. In developing a model, the choice of features comes from somewhere, from someone's (mental) model. Someday perhaps, the elicitation of feature-based causal rules can become part of best practice in engineering ML models.

Causality is also a useful lens through which to view the global/local explanation dichotomy. Type causality is a repeatable pattern of causality; it has the predictive power sought in scientific investigation. Conversely, actual causality is about an occurrence and is analyzed to impute blame [10]. There are bodies of knowledge around both type and actual causality; it may be possible to leverage this in strengthening ML's global/local distinction. These are among perhaps many reasons to dig deeper into causality.

2.3 Systems Engineering

Appeal to ideas in conventional systems engineering (SE). Here I have in mind specifically the matter of (1) establishing system development best practice and workflows, (2) verification and validation (V&V), and (3) failure modes and effects analysis (FMEA). There is no reason to reinvent the wheel when describing a best practice for engineering AI systems; there is much commonality with the SE of other systems. In this regard, your fourth

principle, about competency in the intended operational environment, isn't something unique to ML systems; it is, in fact, a problematic requirement[2] of every engineered system.

I mention FMEA here because, among the various SE techniques, it is easy to see its value to system development best practice. For example, consider the failure modes associated with the system's "explanation capability", they mirror the four principles. The consequences of not providing an explanation, or not providing an explanation effective for the stakeholders, can be traced to effect outside the system boundary. These would be different, of course, for different system contexts (healthcare diagnostics vs. creditworthiness) and for different stakeholders. My point here is that the development of guidance on the best practices of explanation requires systems thinking.

References

[1] James Woodward. *Making Things Happen*. Oxford University Press, Oxford, 2003.

[2] Carl G. Hempel. *Aspects of Scientific Explanation and other Essays in the Philosophy of Science*. The Free Press, New York, 1965.

[3] James Woodward. Scientific Explanation, The Stanford Encyclopedia of Philosophy. In Edward N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford University, 2014.

[4] William Bechtel. Generalizing Mechanistic Explanations Using Graph-Theoretic Representations. In Pierre-Alain Braillard and Christophe Malaterre, editors, *Explanation in Biology*, volume 11, pages 199–225. Springer, 2015.

[5] Peter Machamer, Lindley Darden, and Carl F Craver. Thinking about Mechanisms. *Philosophy of Science*, 67(1):1–25, 2000.

[6] Judea Pearl. An introduction to causal inference. *The international journal of biostatistics*, 6(2):Article 7, 2010.

[2] Perhaps you chose "principle" over "requirement" intentionally. I think all four of the principles you cite are more accurately viewed as archetypes for requirements. However, I understand why, this early in the investigation, you would not want to claim that you are defining requirements of the best practice.

[7] Ronald N. Giere. *Explaining Science*. University of Chicago Press, Chicago, 2010.

[8] Kareem Khalifa. *Understanding, Explanation, and Scientific Knowledge*. Cambridge University Press, Cambridge, 2017.

[9] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[10] Y. Halpern, Joseph. *Actual Causality*. The MIT Press, Cambridge, MA, 2016.