BNH.AI

November 17, 2020

National Institute of Standards and Technology
Information Technology Laboratory
100 Bureau Drive
Gaithersburg, MD 20899

Subject: Comments on NIST *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence* Draft Whitepaper

To Whom It May Concern:

Thank you for the opportunity to review the intriguing and useful draft whitepaper *Psychological Foundations of Explainability and Interpretability in Artificial Intelligence*. Below, we present several actionable points of feedback that we believe would improve the final whitepaper.

1. **Provide succinct definitions for *interpretation* and *explanation*.**

The draft suggests:

- Interpretation is a concept relating to the mechanisms or outputs of a machine learning system that is rooted in background knowledge and goals and allows for high-level decision making. Interpretation is often presented in response to questions of *why*, and interpretation corresponds with the notion of "gist" from fuzzy-trace theory.

- Explanation is a concept relating to the mechanisms or outputs of a machine learning system that is technical and causative. Explanations are typically more appropriate for technical practitioners and enable debugging tasks. Explanation is often presented in response to questions of *how*, and explanation corresponds with the notion of "verbatim" from fuzzy-trace theory.

Succinctly and explicitly summarizing these concepts, as above, would be highly useful for both regulators and practitioners.

***Disclaimer**: bnh.ai leverages a unique blend of legal and technical expertise to protect and advance clients' data, analytics, and AI investments. Not all firm personnel, including named partners, are authorized to practice law.*

2. Bolster design recommendations using currently existing interpretable models, post-hoc explanations, and interactive graphical interfaces.

Given the rich set of currently existing white-box and grey-box models, post-hoc explanations techniques, and user-interaction modalities, it may be possible to offer more concrete recommendations than those proposed throughout the draft, and particularly in Section 4.1.

For example, it seems that techniques such as Shapley values, gradient-based feature attributions, and other locally accurate post-hoc explanation techniques correspond roughly with the concept of verbatim explanations in the draft. To informed practitioners, these types of explanations can rise to the level of a causal explanation.

Being interpretable, as defined in the draft, is a lofty goal for currently existing machine learning transparency techniques. However, it does seem possible that pairing (a) so-called interpretable machine learning models, such as explainable neural networks, scalable Bayesian rule lists, or monotonically constrained gradient-boosting machines, with (b) purpose-built graphical user interfaces (GUIs) such as Google's What-If tool, could bring those transparency techniques in line with the level of interpretability envisioned in the draft.

Referencing currently existing transparency tools, as opposed to general machine learning models and concepts[1], would enable more specific and actionable guidance for regulators and practitioners. More actionable takeaways based on newer technologies could include:

- Using both low-level post-hoc explanations and high-level interpretable models paired with purpose-built GUIs is most likely to provide the necessary verbatim and gist information required for different stakeholder personalities to reason about model mechanisms and outcomes.

- Using low-level, high-fidelity and high-level, low-fidelity post-hoc explanations paired with purpose-built GUIs is most likely to provide the necessary verbatim and gist information required for different stakeholder personalities to reason about model mechanisms and outcomes.

- Contextualizing interpretations for high-level decision makers in terms of background knowledge and goals is most effective.

[1] *E.g., logistic regression models, deep learning, and support vector machines (SVMs).*

- Tailoring interpretations and explanations, and associated accuracy measurements, to the personality types of stakeholders is most effective.

3. **Give more consideration to currently existing notions of interpretability and explainability in machine learning.**

The draft presents compelling and new definitions for interpretability and explainability that are rooted in human perceptions of machine learning. For maximum impact, consider discussing the current notions of these topics, which are often rooted in trade-jargon, and justify why your proposed definitions are (a) connected to current definitions and (b) superior.

Some current machine learning definitions of interpretability and explainability to consider include:

- **Adverse Action Notices (AANs).** These are defined by the Equal Credit Opportunity Act, and its implementation in Regulation B, and the Fair Credit Reporting Act. AANs are a long-standing method of predictive modeling decision interpretations for consumers in the United States.

- **Interpretable.** A relation between a model M, a user U, and a query Q. If the user can answer the query by inspecting the model, then M is interpretable for that (U,Q) pair. Given two models M_1 and M_2, M_1 is more interpretable than M_2 for (U,Q) if it is easier for U to answer Q by inspecting M_1 than by inspecting M_2.[2] In plain English, it has also been defined as "the ability to explain or to present in understandable terms to a human."[3]

- **Explanation.** A collection of visual and/or interactive artifacts that provide a user with sufficient description of the model behavior to accurately perform tasks like evaluation, trusting, predicting, or improving the model.[4]

- **Interpretable Machine Learning**: A model that is "constrained in model form so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge."[5]

[2] *See Dietterich, https://twitter.com/tdietterich/status/1053647527873077251.*

[3] *See Doshi-Velez & Kim, https://arxiv.org/pdf/1702.08608.pdf.*

[4] *See Hall et al. (quoting Sameer Singh), https://arxiv.org/pdf/1906.03533.pdf.*

[5] *See Rudin, https://arxiv.org/pdf/1811.10154.pdf.*

> **Explainable Machine Learning:** "A second (post[-]hoc) [technique] [that] is created to explain the first black box model.[6]

As expressed in (2), it also appears there is some linkage between current notions of low-level, locally-accurate post-hoc explanations and the draft's proposed definition of explanation. It also appears that there is a linkage between the draft's proposed definition of interpretation and high-level, low-fidelity post-hoc explanations, so-called interpretable models, and human-centered GUIs purpose-built for increased transparency in machine learning. Again, highlighting how the proposed definitions build on these currently-existing notions could also make the draft more impactful.

4. Include visualizations of currently existing explanation and interpretation technologies as figures in the text.

It would be illuminating to include visualizations of current explanation and interpretation techniques as figures in the draft. A figure as simple as a plot of Shapley values or a Grad-CAM illustration could represent explanation. Something like the Caruana study in which a so-called interpretable model, a GA2M, was used to help physicians make medical decisions and discover logical flaws in rule-based systems, could rise to the level of interpretation.

5. Include citations to seminal explainable AI (XAI) scholarship.

Line 109 cites "highly-cited literature," but does not include seminal papers published on XAI:

- Ribeiro et al., *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, https://dl.acm.org/doi/abs/10.1145/2939672.2939778. This paper has 4,131 citations.
- Lundberg & Lee, *A Unified Approach to Interpreting Model Predictions*, http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predicti. This paper has 1,918 citations.

[6] *Id. Note that the original quote is technically incorrect, since sometimes the explanation is not another "model."*

6. **Reference existing literature that addresses key issues in the draft.**

- Line 115 asks "interpretable for whom?" However, the draft does not cite the Tomsett et al. article with the same title: *Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems*.[7]

- The Poursabzi-Sangdeh et al. paper, *Manipulating and Measuring Model Interpretability*,[8] appears credible and relevant to the subject of the draft, but does not appear to be cited anywhere.

7. **Address privacy and discrimination risks in the rental and medical diagnosis examples.**

Property rental and medical diagnosis are both applications of machine learning that have received negative attention recently. Given that enhanced transparency is a major mitigating factor in algorithmic discrimination and data privacy incidents, it may be approrpriate to touch on the need for, and benefits of, increased transparency for machine learning in these applications. Introducing recent criticisms and incidents may serve to motivate the examples more clearly. These include:

- *Landlord Tech Watch*: https://antievictionmappingproject.github.io/landlordtech.

- *The Locked Out Series*: https://themarkup.org/series/locked-out.

- *How an Algorithm Blocked Kidney Transplants to Black Patients*: https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients.

8. **Clarify that machine learning models do generate multiple representations of concepts.**

The sentence at line 338 states that "algorithms typically generate a single representation, or model, of a dataset." This sentence could be perceived as problematic for several reasons, including:

[7] *Available at https://arxiv.org/pdf/1806.07552.pdf.*
[8] *Available at https://arxiv.org/pdf/1802.07810.pdf.*

- In reality, every prediction is the expected value of some distribution of predictions. Relatedly, SHAP values account for this by averaging across a distribution of predictions for a single row of data to generate a *consistent* local explanation.

- By definition, deep learning models learn hierarchical sets of representations.

- There are multitask learning algorithms.

9. **Introduce local feature importance in Section 2.2.**

While many authors emphasize contrastive explanations as a better paradigm for consumers, the local feature importance paradigm, which is not contrastive, is probably the most popular way for practitioners to interact with technical explanations. And among locally accurate feature importance techniques, those that sum to the prediction value for a given row of data are often perceived as more trustworthy.

10. **Mention accurate, interpretable models in the paragraph starting at line 656.**

The advent of accurate and interpretable models, such as GA2M, seems to fit nicely with the discussion of coherence and correspondence. Interpretable, accurate models like GA2M are also likely to be important technical drivers of interpretation described in the draft.

11. **Synchronize language in Section 3.1.1 with the *Four Principles* draft terminology.**

The "white box," "grey box," and "black box" language is inconsistent with the recently released *Four Principles* draft terminology. However, "white box" is a much more recognizable term than "self-explaining model." We recommend changing the *Four Principles* draft to use "white box" or "interpretable" over "self-explaining," thus keeping all terminology in sync.

12. **Tie Sections 3.2.1 and 3.3 to overarching normative goals.**

We recommend more clearly tying Section 3.2.1 to improving transparency in machine learning. In the alternative, you could explicitly state why the section is important. As it currently stands, the section seems out of place.

Additionally, there seems to be a direct link between the goals of systems engineering, as described in Section 3.3, and the broader notion of responsible AI. A primary tenant of responsible AI is to bring together diverse expertise to design systems that are aware of many different risks, and account for these risks during system design. We recommend explicitly tying the two concepts together.

13. Fix minor typographical errors.

These include:

- **Line 427:** models? Or model outputs?
- **Line 456:** inconsistent capitalization of machine learning then, line 518, abbreviated?
- **Line 545:** duplicative "in"
- **Line 566:** missing "and"
- **Line 654:** "disclosed" should be "disclose"
- **Line 691:** replace "regressions" with "linear models" (minor technical error)[9]
- **Line 816:** missing closing parenthesis

We hope you find these comments helpful and actionable as you prepare the final document. We are happy to provide additional clarification and discussion around any of the provided feedback, and to be helpful as is appropriate going forward.

In the meantime, thank you for your efforts in this difficult, yet critical, area.

Andrew Burt
Managing Partner, bnh.ai

Patrick Hall
Principal Scientist, bnh.ai

[9] *Regression, i.e., predicting an interval numeric value, can be done with a black-box model.*