BNH
.AI

October 1, 2020

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

Subject: Comments on NIST XAI Draft Whitepaper

To Whom It May Concern:

On our reading of the National Institute of Standards and Technology (NIST) draft whitepaper on explainable AI (XAI), we were impressed by the consideration of different scientific perspectives and the discussion of the many human factors that are inextricably linked to XAI. However, we feel we can provide a few actionable points of feedback, as outlined below.

1. Consider More Careful Treatment of Regulation and Consumer Protections.

Today, laws like the Equal Credit Opportunity Act (ECOA) and the Fair Credit Reporting Act (FCRA) require financial institutions to provide consumers with adverse action notices for negative credit decisions and other types of potentially harmful financial decisions. Because many such decisions are made using predictive models, regulatory commentary provides further clarification on reporting adverse action notices based on predictive models to consumers.[1] In our opinion, it is likely that the appeal and override processes enabled by XAI and adverse action notices will not only be a primary

[1] See, for example: Comment for 1002.9 - Notifications, Paragraph 9(b)(2). For more recent Consumer Financial Protection Bureau (CFPB) comments on explanations for AI systems, see: Innovation spotlight: Providing adverse action notices when using AI/ML models.

***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.*

application of both global and per-decision XAI techniques, but that decision-appeal is a fundamental justification for XAI that could be treated more thoroughly in the draft document.

Furthermore, the current Federal Reserve SR 11-7 Guidance on Model Risk Management also highlights the need for rigorous documentation of predictive models.[2] XAI techniques, such as partial dependence and variable importance plots, are quickly becoming a mainstay of these model risk-management documents, making mandated or voluntary model documentation another key application of XAI techniques.

Among other U.S. government entities, the Federal Trade Commission (FTC) and the Financial Industry Regulatory Authority (FINRA) recently released documents telegraphing increased regulatory oversight for AI systems in the U.S.[3] Given that both the FTC and FINRA highlight the need for transparency and model governance, it is likely that legal requirements around adverse action notices and model documentation could expand to affect a much wider swath of the U.S. economy. Outside of the U.S., organizations like the U.K. Information Commissioner's Office (ICO) and the Singapore Personal Data Protection Committee (PDPC) have put forward highly detailed model audit materials that also suggest accountability, documentation, explanation, and transparency are necessary for AI systems. In fact, the U.K. ICO mentions lawfulness as a major consideration for AI systems in general. The trend of mandated explanation and documentation is therefore becoming visible both in the U.S. and internationally. It may be helpful to reference these additional international documents at line 702.[4]

Acknowledging and providing more thorough legal analysis and reasoning for this trend in the NIST document could help commercial practitioners prepare for the likely future increases in regulatory oversight. Indeed, there are a few places in the existing document that reference existing and future regulations as one major motivation for the current study; leaving this critical area under-analyzed leaves a gap in the document.

Additionally, many of the above referenced regulations mandate some form of explainability—not only to foster trust, as the document describes, but to *empower* and to form the basis of *appeals* for consumers and the general public. Again, further legal analysis and grounding would greatly bolster the document's substantive contributions.

[2] See: Letter SR 11-7 Attachment Supervisory Guidance On Model Risk Management.

[3] See: Using Artificial Intelligence and Algorithms and *Artificial Intelligence (AI) in the Securities Industry*.

[4] See, from the U.K. ICO, for example: What Do We Need To Do To Ensure Lawfulness, Fairness, and Transparency in AI Systems?. See, from the Signapore PDPC: *Model Artificial Intelligence Governance Framework Second Edition*.

2. Provide Suggestions for Measuring Explanation Accuracy.

In our opinion, the XAI practitioner community is hungry for U.S. government organizations, such as NIST and other regulatory agencies, to provide basic guidance on the suitability of XAI approaches. The *explanation accuracy* principle comes tantalizingly close, but does not provide concrete information on assessing explanation accuracy. In conversations with major credit lenders, among other commercial organizations, it is our experience that official guidance on measuring explanation accuracy is needed. In general, well-meaning organizations are worried that they will invest years of work and millions of dollars to explain their ML-based decisions, only to find that the explanation technique they chose is not acceptable to regulators, is inaccurate, or both.

Given all this, it would be extremely helpful for NIST to put forward some ways to measure explanation accuracy. While the draft report does cite human evaluation studies as a potential solution, as proposed by Doshi-Velez, Kim, and others, such studies are potentially time-consuming and difficult for smaller organizations to manage. We are aware of a few papers that attempt to address explanation accuracy from a technical standpoint that were not cited in the draft report. They include:

- *Towards Robust Interpretability with Self-Explaining Neural Networks*, available at https://papers.nips.cc/paper/8003-towards-robust-interpretability-with-self-explaining-neural-networks.pdf
- *Quantifying Model Complexity via Functional Decomposition for Better Post-Hoc Interpretability*, available at https://arxiv.org/pdf/1904.03867.pdf
- *Assessing the Local Interpretability of Machine Learning Models*, available at https://arxiv.org/pdf/1902.03501.pdf

In our own practical work, we have used a handful of empirical strategies to validate explanation accuracy. We have found these strategies helpful, especially when combined with broader model documentation efforts. They include:

- ***Simulated Data***. It is possible to simulate data where global and local drivers of model outcomes are known. Such data can be used to verify XAI techniques.
- ***Comparison to Existing Explanations***. In some cases, previous explanations from logistic regression or other more transparent models might be known. These can provide a baseline to measure against.

- ***Stability Testing and Consistency***. Stability and consistency are oftentimes the biggest problems we see with XAI techniques, aside from Shapley values. For many XAI approaches, a minor change to a predictive model hyperparameter changes all of the local explanations.[5] Testing explanation stability under data or model perturbation can help uncover unstable and inconsistent XAI approaches.

Alternatively, if NIST is of the opinion that human studies are the best current evaluation approach for explanation accuracy, saying so would be helpful.

3. Provide More Treatment of XAI Security and Privacy Risks.

Put bluntly, surrogate models, an XAI technique, enable several adversarial attacks against machine learning (ML) models. Moreover, providing explanations with model predictions makes these attacks easier and increases the potential for privacy harms. These harms include violating the confidentiality of model mechanisms and sensitive training data. We are aware of a number of papers that discuss these subjects but were not cited in the draft report. They include:

- *On the Privacy Risks of Model Explanations*, available at https://arxiv.org/pdf/1907.00164.pdf
- *Membership Inference Attacks Against Machine Learning Models*, available at https://arxiv.org/pdf/1610.05820.pdf
- *Stealing Machine Learning Models via Prediction APIs*, available at https://arxiv.org/pdf/1609.02943.pdf
- *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, available at https://dl.acm.org/doi/pdf/10.1145/2810103.2813677
- *Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers*, available at https://arxiv.org/pdf/1306.4447.pdf

We did see some references to attacking explanations in the current draft, which is a significant point that should remain in the final document. Because attacks on explanation require an even higher degree of understanding of AI systems, we would argue that hacking or stealing data and models directly via XAI approaches is a more realistic attack vector in today's AI systems.

[5] Molnar defines consistency in *Interpretable Machine Learning*, Section 2.5. In our experience, consistency is extremely important for trustworthiness. Especially in the consumer credit world, where explanations are analyzed in the context of adverse action notices, per-decision explanations for a single consumer that change dramatically based on small changes to model or dataset specifications are likely to raise eyebrows. Consistent explanations should be more robust to such changes.

4. Additional Comments.

The following feedback addresses more minor or technical points in the draft document.

- ***Switch "Self-Explainable Models" for "Interpretable Models"*****:** We did not find the phrase "Self-Explainable Models" helpful. The technical practitioner and research communities seem to be consolidating on "interpretable models" as a name for these functions. The addition of a new phrase to describe them only muddies waters that were just beginning to clear. Keith O'Rourke provides several references to bolster our claim, some of which you already consider in the draft document.[6]

- ***Consider Off-label Use a Violation of Knowledge Limits*****:** We suggest mentioning the idea of off-label use, i.e., when a model is intentionally or unintentionally used for purposes that violate its mathematical or business assumptions, being a violation of knowledge limits. As AI systems are packaged as reusable products through prediction APIs and other technologies, they will become easier to use for purposes for which they were not designed. As with myriad other technologies, from prescription medicines to power tools, we anticipate that off-label use will cause harm to both operators and consumers of AI.

- ***Clarify "Types" of Explanation versus "Applications" or "Goals" for Explanation*****:** We suggest changing the title of Section 3, "Types of Explanations," to "Types of Explanation Applications," or "Goals of Explanations." While we found the discussion in Section 3 accurate and informative, we found the title of the section to be slightly misleading. For instance, "Societal Acceptance" does not appear to be a typology for an explanation. It seems much more clear that societal explanation is a goal of, or an application of, explanation. How can we place a technique such as SHAP or partial dependence into one of the buckets defined in Section 3? It seems these techniques would fall into *all* of the types put forth in Section 3.

- ***Remove the Sentence Beginning at Line 388*****:** From a research perspective, per-decision explanation techniques like LIME and SHAP made a huge splash precisely because they were more difficult to formulate and test than global explanation techniques. Global

[6] See: Explainable ML Versus Interpretable ML.

explanation techniques like partial dependence, decision tree surrogate models, and global variable importance had been recognized for decades before the advent of LIME, SHAP, and other highly impactful per-decision explanation techniques. We would therefore suggest striking the sentence that begins on line 388: “Furthermore, global explanations are harder to generate than per-decision explanations because per-decision explanations only require an understanding of a single decision.” This sentence may generate unnecessary controversy without making a material contribution to the document.

- ***Bolster Claims at Line 419***: The claims in line 419 can be bolstered, especially with regards to structured data, by the documented results for GA2M / Explainable Boosting Machine (EBM).[7] Note that these results come with code for reproduction, and that the capability for a model to be more accurate and more transparent than a traditional ML black-box is an extremely important scientific development.

- ***Add More Practical Examples to Section 5.1***: Section 5.1 could include more interpretable models that are used in practice, including constrained variants of traditional ML black-boxes such as monotonic gradient boosting machines (GBM), explainable neural networks (XNN), monotonic neural networks, and Equifax’s NeuroDecision.[8]

- ***Correct a Possible Error in Section 5.2***: In Section 5.2, the SHAP method appears to be artificially linked to regression tasks. SHAP is valid for regression, binomial classification, and multinomial classification tasks. Also, we feel that introducing SHAP as a global explanation technique is misleading. By definition, SHAP is calculated per-decision and aggregated to a global viewpoint. (Of course, SHAP is an excellent global explanation technique, and the visualizations in the Python shap package are perhaps some of the most nuanced ways to present global feature importance.)

- ***Include ALE in Section 5.2***: Partial dependence is one of the oldest global explanation techniques for ML. Thus, it suffers from weaknesses that newer techniques have been invented to address. In particular, we feel it would be important to mention accumulated local effect (ALE) plots in addition to the discussion of partial dependence plots in

[7] See: Interpret project GitHub page.

[8] For monotonic GBM and XNN, see: *A Responsible Machine Learning Workflow*; for monotonic neural networks, see: TensorFlow Lattice; and for NeuroDecision, see: Equifax Receives Utility Patent for Innovative NeuroDecision® Technology.

Section 5.2. ALE was designed to overcome the computational challenges of partial dependence and the problems with partial dependence in the presence of correlated features.[9]

- ***Address SHAP and Local Feature Importance Directly in Section 5.3*****:** Although far from perfect, the local feature importance paradigm, wherein per-decision feature attributions are ranked or plotted as numeric values, is likely the most common way that practitioners interact with per-decision XAI techniques today. Also, SHAP values, due to their theoretical advantages and implementation in popular software like XGBoost, h2o, and lightgbm, are one of the most common approaches to calculate local feature importance. We strongly recommend that these subjects be addressed in Section 5.3 of the draft document.

- ***Frame Section 6 in Terms of Enhancing XAI*****:** As XAI is completely dependent on human comprehension and perception of machine-generated explanations, we found the discussion in Section 6 valuable and fascinating. Furthermore, consideration of these human factors is often missing from open-source and commercial efforts in XAI implementation, resulting in suboptimal XAI products for both practitioners and consumers. This is not simply a design problem. Misinterpretation of ML explanations for high-stakes implementations could lead to full-blown AI incidents. However, we felt the current discussion in the draft document is somewhat detached from the goals of XAI. In our opinion, framing Section 6 in terms of furthering the goals of XAI would strengthen the draft document.

- ***Include Additional Fundamental Machine Learning Concepts*****.** We felt that the draft document downplayed two fundamental technical concepts that are important for interpretable models and XAI:

  - ***The Rashomon Effect*****:** Also known as “the multiplicity of good models,” this concept states that many, many potential ML models exist for any given dataset. The Rashomon effect is important for at least two reasons. First, in all of those models, it is potentially possible to find accurate, stable, and interpretable models.[10] Second, many less-rigorous XAI approaches ignore the Rashomon effect,

[9] See: *Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models*.

[10] See: *A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning*.

causing them to have low consistency—a problem somewhat akin to poor reproducibility. For these techniques, when a hyperparameter of a model changes, the per-decision explanations change. This is not acceptable when an XAI user expects that XAI techniques are explaining phenomenon in a dataset. More rigorous XAI techniques, primarily SHAP, do consider large sets of perturbed models or data, in an attempt to find consistent explanations across sets of similar models.[11] In our opinion, these types of explanations are more likely to be accurate on unseen data and to exhibit better stability when models are refreshed or retrained.

- ***Causality*:** If actual causal factors, instead of noisy, high-degree local dependencies, are the drivers of model predictions, this can greatly enhance model stability and interpretability. It appears that the draft document does not mention this important connection between causal inference and interpretability. Causal modeling approaches could also be mentioned in Section 5.1 as additional types of interpretable models.

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

[11] For an excellent discussion of the "true to data" vs. "true to model" philosophies, and a deep dive on important subtleties of Shapley values in machine learning, see: *True to the Model or True to the Data?*.