NISTIR 832 Review

[Link to NSTIR832 document]

- In Section 4, **Explanation** (as a pillar) has some nuance thatis missed:
    - Explanation Quality
        - Accuracy of explanations, in particular, Shapley Value estimations (see Section V of [Datta, Sen, Zick 2016]). See Sundararajan & Najmi 2020 for additional nuances, including the fact that certain instances of the SHAP library does not even estimate Shapley Values.
        - Capturing causal influence in local explanations, i.e. identifying features that are truly driving the model's predictions and teasing them apart from associated features.
        - The "meaningful" pillar should **incorporate a notion of sufficiency**. Explanations must be understandable but also sensical and enough to justify the predicted outcome. We must leverage important input and internal factors as a way to evaluate sufficiency of explanations [Leino et al 2018, Wang et al. 2020, Lu et al. ACL 2020].
    - Privacy-preserving explanations:
        - Explanations must also ideally retain *privacy*, in that we do not disclose sensitive information about individuals when justifying a prediction. See [Datta, Sen, Zick 2016] for privacy-preserving explanations for Shapley Value feature importances. This is also a challenge for counterfactual explanations and actionable recourse [Karimi et al 2020].
- The provided **definition of global explanations is too narrow**-- it is more than the ability to produce a model that explains/approximates the underlying model. Global explanations like the examples cited in the paper (SHAP, TCAV, ICE, PDPs) and also other work (see here) provide general visualizations or metrics that characterize model drivers overall or in segments.
    - A lot of the same quality requirements that are necessary for per-decision examples also apply global explanations such as: 1) being causally relevant to the behavior of the model, 2) providing per feature explanations.
    - It is important for global and per-decision interpretability methods to be consistent. For example, it should not be the case that a feature that is globally important is unimportant for any decision in particular.
- **Stability as a core component:** In the intro (and beyond) they define AI as "resilient" which is more of a security concern (e.g. resilient to adversarial attacks). Another core component which is ignored is "stability"-- understanding that people and data changes and models must be robust to this or change as well. [SR-11-7]

- While stability might be well-defined from a computer science and statistics perspective, it also plays a role in the psychology of how people trust models.

- Non black-box models like neural networks will **often do better than their whitebox counterparts**. This isn't brought up in Section 5. It seems unfair to say that whitebox models are ideal for trustworthy AI-- while they might be more *interpretable*, they also might have poor *knowledge limits.* As an example, deep learning models are high-dimensional functions and can capture nuances about data that could actually *prevent* bias or make the model decision more robust. It's not just accuracy vs. interpretability but accuracy vs. fairness vs. stability vs. interpretability etc.
    - In Section 5.1, Shapley values are a case where you can get accurate model explanations even if the model is a blackbox.
    - There is a gradient of black box -> white box models. As an example, In Section 5.1, the **claim that GA2M is a whitebox model** is not entirely truthful. While they contain pairwise interactions that are essentially heatmaps, it's hard to interpret the actual feature value of these interactions.
- Would suggest adding information on **calibration** (w.r.t. the metacognition point). In section 6.4, the point they bring up on metacognition has been studied in the statistics and machine learning literature as calibration [Niculescu-Mizil & Caruana, Jiang et al. etc].
- Would suggest adding information on **actionable recourse** within the counterfactual explanations section-- the benefit of counterfactuals are that you are stress-testing models on modified data points. This ties in directly to the psychology of using trustworthy AI-- playing out what-if scenarios, and allowing laypeople to understand how decisions can be changed. Some citations: [Rawal & Lakkaraju, 2020; Karimi et. al 2020; Poyiadzi et. al 2020, etc.]
- In Section 6.2 and Section 6.3, it is not enough for humans to be able to use model explanations to justify a model decision. Instead, if given access to model explanations, humans should be able to **replicate the model decision**-- i.e. come to the same conclusion on their own. This prevents humans from implicitly trusting a model and then retroactively trying to justify it, which is a major problem that is discussed. This ties directly into the **sufficiency of explanations** mentioned earlier; see [Leino et al 2018, Wang et al. CVPR Workshop 2020, Lu et al. ACL 2020].
- We'd like to see **bias and fairness** mentioned in 6.4. Knowing your "knowledge limits" also means understanding what conscious/unconscious biases you possess and trying to actively undo that when making a decision. Humans do this all the time, and models need to do the same. 6.4

- This is not as simple as the examples they gave of model confidence/identifying out-of-distribution points. You could (confidently) reflect human biases in a model even with abundant data.
- We as humans have predefined protected classes that you cannot discriminate on. Quantifying disparate impact should be a prerequisite to adoption of any AI. [Datta, Sen, Zick 2016; Feldman et al 2016; Dutta et al 2020]

References

Dutta, Sanghamitra, et al. "Fairness Under Feature Exemptions: Counterfactual and Observational Measures." *ArXiv:2006.07986 [Cs, Math, Stat]*, June 2020. *arXiv.org*, http://arxiv.org/abs/2006.07986.

Feldman, Michael, et al. "Certifying and Removing Disparate Impact." *ArXiv:1412.3756 [Cs, Stat]*, July 2015. *arXiv.org*, http://arxiv.org/abs/1412.3756.

Karimi, Amir-Hossein, et al. "A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects." *ArXiv:2010.04050 [Cs, Stat]*, Oct. 2020. *arXiv.org*, http://arxiv.org/abs/2010.04050.

Leino, Klas, et al. "Influence-Directed Explanations for Deep Convolutional Networks." *ArXiv:1802.03788 [Cs, Stat]*, Nov. 2018. *arXiv.org*, http://arxiv.org/abs/1802.03788.

Lu, Kaiji, et al. "Influence Paths for Characterizing Subject-Verb Number Agreement in LSTM Language Models." *ArXiv:2005.01190 [Cs]*, May 2020. *arXiv.org*, http://arxiv.org/abs/2005.01190.

Poyiadzi, Rafael, et al. "FACE: Feasible and Actionable Counterfactual Explanations." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Feb. 2020, pp. 344–50. *arXiv.org*, doi:10.1145/3375627.3375850.

Rawal, Kaivalya, and Himabindu Lakkaraju. "Interpretable and Interactive Summaries of Actionable Recourses." *ArXiv:2009.07165 [Cs, Stat]*, Sept. 2020. *arXiv.org*, http://arxiv.org/abs/2009.07165.

Sundararajan, Mukund, and Amir Najmi. "The Many Shapley Values for Model Explanation." *ArXiv:1908.08474 [Cs, Econ]*, Feb. 2020. *arXiv.org*, http://arxiv.org/abs/1908.08474.

Wang, Zifan, et al. “Interpreting Interpretations: Organizing Attribution Methods by Criteria.” *ArXiv:2002.07985 [Cs]*, Apr. 2020. *arXiv.org*, http://arxiv.org/abs/2002.07985.