Harnessing Explanations to Bridge AI and Humans

Vivian Lai
University of Colorado Boulder
Boulder, CO, USA
vivian.lai@colorado.edu

Samuel Carton
University of Colorado Boulder
Boulder, CO, USA
samuel.carton@gmail.com

Chenhao Tan
University of Colorado Boulder
Boulder, CO, USA
chenhao@chenhaot.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Machine learning models are increasingly integrated into societally critical applications such as recidivism prediction and medical diagnosis, thanks to their superior predictive power. In these applications, however, full automation is often not desired due to ethical and legal concerns. The research community has thus ventured into developing interpretable methods that explain machine predictions. While these explanations are meant to assist humans in understanding machine predictions and thereby allowing humans to make better decisions, this hypothesis is not supported in many recent studies. To improve human decision-making with AI assistance, we propose future directions for closing the gap between the efficacy of explanations and improvement in human performance.

Why Do We Need Explanations?
Recent trends in machine learning have led to models that are increasingly powerful, complex, opaque, and ubiquitous. Model performance has begun to meet or exceed expert human performance in numerous areas such as recidivism prediction [8] and medical diagnosis [1]. Concomitantly, AI models have begun to play a larger and larger role in aspects of life such as government, business, and science, leading to ever-higher consequences for model mistakes.

Unfortunately, while average model performance has approached human levels, models still lag behind humans in key ways. AI models tend to absorb bias from their training data, are vulnerable to adversarial inputs, and have difficulty generalizing beyond the specific distribution of that training data [6].

A common suggestion to mitigate these issues is to view models as **augmenting** rather than replacing human effort. In the ideal scenario, a human and a model could work together as a hybrid system whose performance would exceed that of either agent operating alone. AI explanations have been proposed as a way to achieve this cooperation. The hypothesis is that if a human can scrutinize the logic behind a model prediction, they can recognize when that prediction is unfair, nonsensical, or otherwise unreliable [6].

The Current State of Explanations

In order to achieve a balance between AI accuracy and human intuition, the research community has proposed a number of techniques for explaining the predictions of AI models. A common approach is feature attribution, which attempts to assign each feature (word, pixel, etc.) a score indicating its importance in the model's prediction. Such methods range from retroactive perturbation-based analysis like the popular LIME [14] to built-in attention mechanisms such as that proposed by Lei et al. (2016) [12].

However, the community has struggled to demonstrate an improvement in human decision quality as a result of these kinds of explanations. Typical experimental design involves human subjects making decisions in the presence of model predictions and evaluating whether explanations improve their accuracy in doing so. Some experiments in this vein have included predicting apartment prices [13], detecting deceptive online reviews [11], assessing social media toxicity [3], performing various artificial tasks [9], and recidivism prediction [5]. We are not aware of any such experiment that has reported a significant improvement in accuracy that cannot be explained by increased subject trust in a model whose accuracy is higher than the human baseline (such as Lai and Tan (2019) [11]).

Why have explanations failed to improve human performance? While this is a difficult question to answer, existing results provide a few clues. First, Lai et al. (2020) point out two distinct types of AI learning problem: *emulating* human skill vs. *discovering* new knowledge [10]. They speculate that in the latter case, humans may not have strong enough task intuitions to make effective use of simple explanations, leading to a need for additional training [10]. Even in emulation tasks, models may incidentally learn patterns that simply do not correspond well with human intuition, as was observed by Feng et al. (2018) in the case of LSTM models for sentiment analysis [4]. Explanations may be better suited for catching certain type of model errors over others: Carton et al. (2020) observe that they reduce false positives while increasing false negatives, surmising that subjects find it easier to overturn phrases incorrectly identified as toxic than to discover truly toxic phrases missed by the model [3].

Overall, these results suggest a fundamental misalignment between AI explanations and human mental models, a situation that Bansal et al. (2019) discuss as a general hurdle in human-AI collaboration [2]. As a solution, we suggest two basic directions for future work: 1) augmenting human mental models to cope with model explanations; and 2) adjusting model explanations and behavior to match human mental models.

Direction I: Augmenting Human Mental Models

Model-driven tutorials. Humans seem to not have strong intuition in making effective use of explanations in tasks that *discover* new knowledge [10]. To improve human mental models, we propose model-driven tutorials that elucidate counter-intuitive and inconspicuous patterns embedded in models learned from the dataset. Model-driven tutorials are one possible way to align human mental models and AI, and we call for more study on how to effectively train humans to work with AI explanations.

Interactive explanations. The goal of interactive explanations is to allow humans to understand the model better through trial-and-error scenarios. As compared to static explanations that only reveal what is important to the model, interactive explanations allow humans to interact with models and explanations, e.g., by editing input and examining the differences in a model's prediction. Instead of simply presenting important patterns in the model, it is useful for humans to identify patterns through active learning.

Evaluating generalization. It is important to point out that a typical setup in prior work employs a random split to obtain training and testing data, which is a standard assumption in supervised machine learning. While humans can ideally improve generalization in this case, humans might be more likely to correct generalization errors in machine learning models when the testing distribution differs from training. In that case, understanding the embedded patterns, especially spotting spurious ones, can help humans generalize these data-driven insights and *reduce model biases*. A significant challenge lies in how we can properly evaluate such generalization, relating to a core issue in machine learning.

Direction II: Towards Human-Centered Explanations

Understanding human explanations. Existing techniques tend to optimize explanations for numeric qualities like sparsity or some notion of fidelity to the model. Ultimately, however, we need to recognize that explanations serve as a communicative device to humans. Key to this idea is more effort to understand the rationales behind human decisions, the qualities of those rationales associated with correct and incorrect decisions, and the effect of model-human rationale alignment on model-human agreement. Studies such as Kaushik et al. (2019) [7] which collect human rationales/explanations are a good start, but we call for a **behavioral** and **design** perspective on such data rather than its use merely as additional training signal.

Experimenting with alternative explanation types. Feature attribution may simply be inadequate for affording meaningful human oversight of model predictions, especially in discovery-type tasks where they don't have strong existing intuitions. Example-based explanations and natural language explanations may succeed where feature-based explanations fail. Therefore, we call for more human subject experimentation involving alternative explanation styles.

Explanations as model criticism. Another focus area we suggest is to break away from treating explanations as a diagnostic signal for the reliability of a static model. Perhaps instead we should treat them as a means for critiquing the underlying logic of model decisions that are known to be incorrect. While the idea of "learning from explanations" has a long history [15], we are not aware of work that employs this idea in a dynamic way, in response to known model errors, and which incorporates existing model explanations.

Conclusion

AI explanations have generated great excitement as a way to provide added value in high-stakes decision-making. However, they have been failing in recent studies to live up to their promise. We suggest new research directions to address this expectation gap, based on the idea of aligning AI and human mental models to enable the type of critical human scrutiny that is likely to lead to real improvements.

REFERENCES

[1] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* (2019).

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. In *AAAI*.

[3] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Attention-Based Explanations Don't Help Humans Detect Misclassifications of Online Toxicity. In *ICWSM*.

[4] Shi Feng, Eric Wallace, II Grissom, Mohit Iyyer, Pedro Rodriguez, Jordan Boyd-Graber, and others. 2018. Pathologies of neural models make interpretations difficult. *arXiv:1804.07781* (2018).

[5] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *ACM* CSCW (2019).

[6] Riccardo Guidotti, Anna Monreale, Franco Turini, and Dino Pedreschi. 2018. A Survey Of Methods For Explaining Black Box Models. *arXiv:1802.01933* (2018).

[7] Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2019. Learning the Difference that Makes a Difference with Counterfactually-Augmented Data. *arXiv:1909.12434 [cs, stat]* (2019).

[8] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2017. Human decisions and machine predictions. *The quarterly journal of economics* (2017).

[9] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *arXiv:1902.00006* (2019).

[10] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *CHI*.

[11] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *FAT**.

[12] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *EMNLP*.

[13] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]* (2018).

[14] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*.

[15] Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *NAACL*.