Question-Driven eXplainable AI: Re-framing the Technical and Design Spaces of XAI

Q. Vera Liao
IBM Research AI
vera.liao@ibm.com

ABSTRACT

A critical component for people to interact with a ML model is its explainability. The research community have produced an abundance of techniques to produce model explanations, some of which are entering data science practitioners' toolbox with recent availability of explainable AI (XAI) open-source toolkits. However, there exists a disconnect between these XAI techniques and user-centered needs for model explanations, not only making it challenging for practitioners to navigate the space of XAI, but also hampering research effort developing XAI techniques effective for real-world AI applications. Drawing on human-computer interaction (HCI) literature and our research studying the design space of XAI, we propose *question-driven XAI*–re-framing the XAI technical space by what kind of user questions each XAI technique can answer. This re-framing will enable practitioners to make user-centered choices to develop XAI and guide the discovery of opportunities for new XAI algorithms to address real-world user needs. Towards this goal we develop an XAI Question Bank, by gathering common user questions across 16 AI applications, and provide a mapping guide between these questions and XAI techniques.

KEYWORDS

explainable AI, HCI, human-centered AI

ACM Reference Format:
Q. Vera Liao. 2021. Question-Driven eXplainable AI: Re-framing the Technical and Design Spaces of XAI. In *Proceedings of DaSH@KDD'21, August 15-16, 2021, Virtual Conference (DASH@KDD '21).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Making AI explainable, i.e., providing appropriate information to help people *understand* its decisions and behaviors, is a necessary foundation to serve diverse needs of stakeholders, from data scientists' model debugging, regulators' model auditing, to helping various kinds of end-users to achieve their goals. These critical needs for explainability have spurred great academic interests in developing algorithms and techniques to produce model explanations, from algorithms to train "glass box" interpretable models to "post-hoc" techniques to generate explanations for "opaque box" models in various forms [7–9, 12, 16]. Recently, open-source toolkits such as IBM AIX 360 [3], Microsoft InterpretML [4], H2o.ai ML Interpretability [1] are making popular XAI algorithms (e.g. [10, 17–19]) accessible to data science practitioners, both to aid model development and provide model explanations to stakeholders.

However, AI practitioners, both data scientists and AI product teams in general, face challenges navigating this growing technical space of XAI. Although some taxonomies of XAI algorithms exist [9, 12], to make appropriate choices of XAI techniques, whether to leverage open-source solutions or implement model-specific solutions, requires first to understand what kind of explanation is needed by the recipient. Such explainability needs are varied and dynamic. While some recent works start to delineate common user groups of XAI [8, 14, 20], even the same user may demand different types of explanation at different usage points, such as during on-boarding stage versus seeing surprising predictions. After understanding users' explainability needs, the challenge is to identify the right XAI techniques that can accommodate the needs, and sometimes practitioners may find such techniques to no avail.

To facilitate practitioners to make user-centered choices of XAI techniques, and to encourage the research community to center the XAI research efforts on real-world user needs, we propose a question-driven framework to contextualize XAI techniques by what kinds of user questions they can answer. The framework has two components: 1) an XAI Question Bank (Figure 1) of common questions users have for understanding ML systems. 2) A mapping between these user questions and XAI techniques (Figure 2).

For the first part, the idea to represent different types of explanations by the types of questions they can answer has its roots in HCI [15] and social sciences literature [13]. We consulted with this prior work to arrive at nine prototypical categories of user questions for explanations in the ML context, including *Why*,*How* (the model works globally), *What if*, etc. (see Figure 1). We took a broad view on explainability–any information to help people better understand the model, and also included three "model facts" types of explanations based on [15]:*data*, *output*, *performance*. We then validated these categories and gathered concrete forms of user questions by "crowd-sourcing" common user questions from 20 designers working across 16 ML products. This was part of a broader study on design practices for XAI, detailed in [blinded for review], where we also discussed high-level design guidelines to address each category of explainability needs, and provided an analysis on where the current landscape of XAI techniques may fall short in addressing these real-world user needs. For example, explanations of data limitations, multiple models, model changes, counterfactual explanations, and personalizing explanation to the right level of granularity, remain challenging but needed.

For the second part, we provide a mapping guide between the user question categories in Figure 1 and XAI techniques. Through

DASH@KDD '21, August 15-16, 2021, Virtual Conference

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

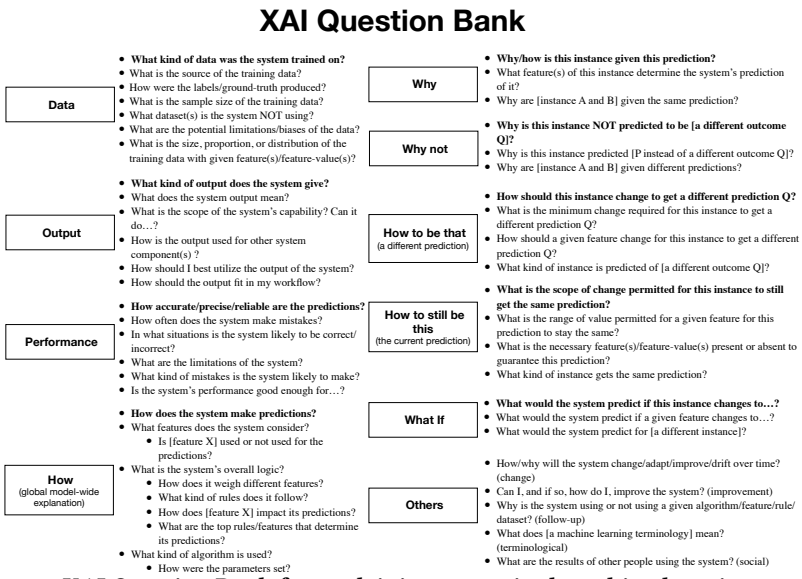

Figure 1: XAI Question Bank for explaining supervised machine learning systems

Question	Explanations	Example XAI techniques
Global how	• Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view • Describe the general model logic as feature impact*, rules+ or decision-trees• (sometimes need to explain with a surrogate simple model)	ProfWeight*+•, Feature Importance*, PDP*, BRCG+ , GLRM+ , Rule List+ , DT Surrogate•
Why	• Describe what key features of the particular instance determine the model's prediction of it* • Describe rules+ that the instance fits to guarantee the prediction • Show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors+, ProtoDash•
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction* • Show prototypical examples+ that had the alternative outcome	CEM* , Prototype counterfactual+ , ProtoDash+ (on alternative class)
How to be that	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction* • Show examples with minimum differences but had a different outcome than the prediction+	CEM*, Counterfactuals*, DiCE+
What if	• Show how the prediction changes corresponding to the inquired change	PDP, ALE, What-if Tool
How to still be this	• Describe feature ranges* or rules+ that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome	CEM*, Anchors+
Performance	• Provide performance metrics of the model • Show confidence information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Confidence FactSheets, Model Cards
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	• Describe the scope of output or system functions • Suggest how the output should be used for downstream tasks or user workflow	FactSheets, Model Cards

Figure 2: Mapping guide between question categories and XAI techniques, with links to code libraries

this mapping we derive guidelines (column 2) for explanations to address each category of user question, based on available techniques. While we focus on popular XAI techniques available in current open-source toolkits [1, 3–5] for practical use, both the question bank and the mapping are extendable. For the last three categories of "model facts", the corresponding mapping (FactsSheets [6], Model Cards [2] and Datasheets [11]) provide exemplars of industry standards for model documentation. This mapping was built by reviewing various survey papers of XAI techniques [7, 8, 12], and iterated on with feedback from 5 XAI researchers.

By re-framing the technical space of XAI with this mapping between XAI techniques and user questions grounded in behavioral theories and empirical research, we encourage practitioners to foreground user needs instead of technical affordance when developing model explainabiltiy features, for example, by first understanding user questions in different interaction scenarios. We further propose a question-driven user-centered XAI design process that brings together data scientists and designers to ideate on explainability features based on user questions, detailed in [blinded for review]. More fundamentally, we believe this re-framing could enable constructing the technical space of XAI in a more user-centered way, encouraging XAI researchers to articulate what user needs are served by the algorithms they develop, and examine the technical gaps and opportunities to address real-world user needs for explainability.

REFERENCES

[1] 2017. H2O.ai Machine Learning Interpretability. https://github.com/h2oai/mli-resources.

[2] 2019. Google Cloud Model Cards. https://modelcards.withgoogle.com/about.

[3] 2019. IBM AIX 360. aix360.mybluemix.net/.

[4] 2019. Microsoft InterpretML. hhttps://github.com/interpretml/interpret.

[5] 2019. SeldonIO Alibi Explain. https://docs.seldon.io/projects/alibi/en/stable/.

[6] 2020. IBM AI FactSheets 360. https://aifs360.mybluemix.net/.

[7] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[9] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[10] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 590–601.

[11] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[12] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[13] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[14] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.

[15] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

[16] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[17] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[20] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).