Towards Useful AI Interpretability via Interactive AI Explanations

Hua Shen
huashen218@psu.edu
Pennsylvania State University
PA, USA

ABSTRACT

Advancements in deep learning have revolutionized AI systems, enabling collaboration between humans and AI to enhance performance in specific tasks. AI explanations play a crucial role in aiding human understanding, control, and improvement of AI systems regarding various criteria such as fairness, safety, and trustworthiness. Despite the proliferation of eXplainable AI (XAI) approaches, the practical usefulness of XAI in human-AI collaborative systems remains underexplored. This doctoral research aims to **evaluate and enhance the usefulness of AI explanations in practical human-AI collaboration**. Five projects were conducted to investigate: *1)* the usefulness of state-of-the-art AI explanations for humans, *2)* the disparities between AI explanations and user demands in practical settings, and *3)* strategies to empower useful AI explanations for human-AI collaborative systems. Our findings indicate that cutting-edge AI explanations are largely not useful, potentially due to the mismatch between diverse user needs and limited XAI displays. Therefore, we propose interactive AI explanations as a means to enable users to inquire about specific explanations with minimal cognitive load. To this end, we developed two interactive AI explanation systems that improved human-perceived usefulness in terms of both human performance and model performance within human-AI collaborative systems. Finally, we discuss the limitations and challenges of achieving useful XAI for future research.

KEYWORDS

human-centered XAI, useful AI explanation, interactive XAI

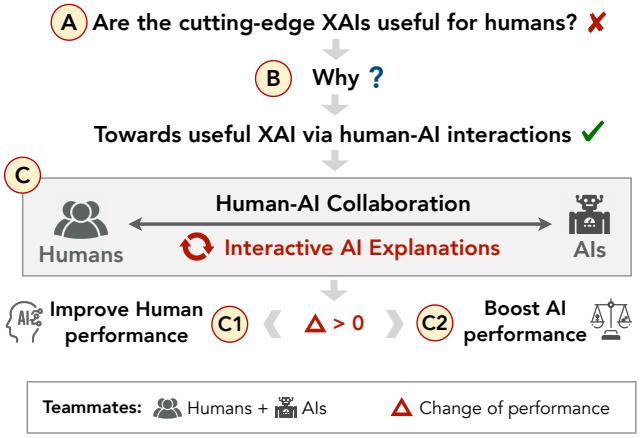


Figure 1: An overview of this doctoral thesis. To evaluate and improve the XAI usefulness, we investigate: (A) are cutting-edge XAI useful for humans? [10, 13] (B) what's the disparity between XAI and practical user needs? [11] and (C) how to empower useful XAI via interactions? Particularly, we develop two interactive AI explanation systems to improve (C1) human [9] and (C2) AI model [17] performance, respectively.

1 INTRODUCTION

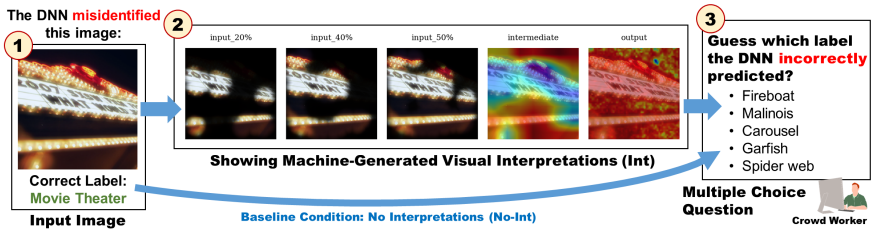
Deep learning advancements have led to breakthroughs in numerous artificial intelligence (AI) systems [8, 14, 15]. Therefore, humans collaborate with the superior capability of AI systems to achieve complementary performance on specific tasks [1, 12]. The complex applications of human-AI collaborative systems have led to a surge of interest in developing systems, which are not only optimized for task performance but also require catering to other vital criteria such as fairness for demographic groups, safety on attacks, human trustworthiness [14, 18], etc. For human-AI collaborative systems, ensuring these auxiliary criteria is of great importance. However, these auxiliary criteria, unlike the conventional task performance metrics (*e.g.,* accuracy), are commonly qualitative measures that are difficult to be quantified. Here is why we need AI interpretability criterion – interpretability per se is not our goal, instead, AI interpretability is a fallback to be used by humans to gauge the AI model reasoning and assess the auxiliary measurements [4].

Therefore, a surge of eXplainable AI (XAI) approaches have been developed and validated to faithfully reflect the model behavior with the automatic metrics such as *faithfulness* [3, 6] and be plausible to humans assessed by the metrics like *plausibility* [2]. However, taking one step further to fulfill the practical XAI roles, it is still under-explored in terms of how humans can leverage AI explanations, as the auxiliary criteria, to boost human-AI collaborative systems in practical applications such as debugging the model or simulating model prediction, etc [5].

In this paper, the **overall objective** is to **evaluate and improve the usefulness of AI explanations** for humans-AI collaborative systems in real-world practice. I break down the research objective into investigating three research questions:

- **RQ1**: Are cutting-edge AI explanations useful for humans in practice? (Section 2)
- **RQ2**: What's the disparity between AI explanation and practical user demands? (Section 3)
- **RQ3**: How to empower useful AI explanation with human-AI interaction? (Section 4)

We examined the three research questions by conducting five projects. To answer RQ1 (Fig 1 (A)), we deployed two real-world human evaluation studies on analyzing computer vision AI model errors with post-hoc explanations [10], and simulating NLP AI model predictions with inherent explanations [13], respectively. The two studies unveil that, surprisingly, AI explanations are not always useful for humans to analyze AI predictions in practice. This motivates our research for RQ2 (Fig 1 (B)) – gaining insights into



4

	C1	C2	C3	C4	C5	Overall
Int	0.57	0.74	0.66	0.41	0.67	0.63
No-Int	0.52	0.71	**0.84	*0.59	0.77	**0.73
#images	44	20	112	18	6	200

Table: The machine-generated visual interpretation again reduced the average human accuracy in inferring model misclassification.
(Paired t-test: *: p<0.05, **:p<0.01).

Figure 2: The workflow of human study 1 [10]: "Guessing the Incorrect AI Predicted Label Task". The humans are presented with an image (1), a set of post-hoc AI explanations (2), and are then asked to guess the incorrectly predicted label from five candidates (3). The results (4) show that displaying AI explanations decrease human average accuracy by roughly 10%.

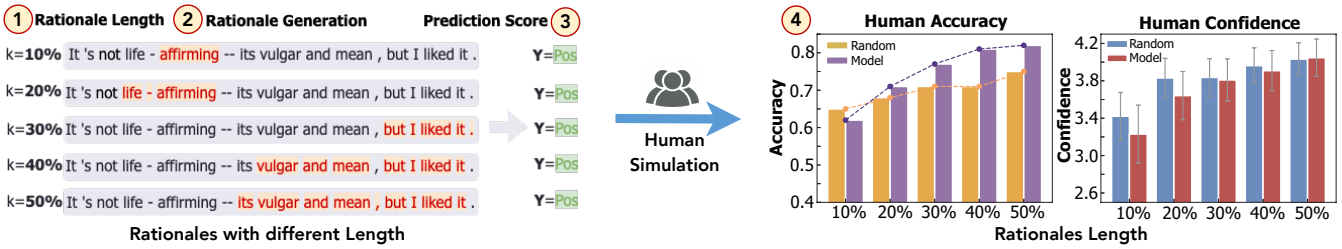


Figure 3: An toy example and result demonstration of human study 2 [13]: "Simulating the Correct AI Prediction Task". Humans are shown with AI self-generated rationales (2) with different lengths (1), and are asked to simulate the AI predictions (3). Results (4) show that shortest rationales are largely not useful for humans to simulate AI text classifiers.

disparities between the status quo of AI explanations and practical user needs [11]. By surveying over 200 AI explanation papers and comparing with summarized real-world user demands [7], we observe two dominating findings: *i)* humans request diverse XAI questions across the AI pipeline to gain a global view of AI system, where existing XAI approaches commonly display a single AI explanation that can not satisfy diverse XAI user needs; *ii)* humans are widely interested in understanding what AI systems can not achieve, which might lead to the need of interactive AI explanations that enable humans to specify the counterfactual predictions.

In light of these findings, we deeply deem that, instead of designating user demands by XAI researchers during AI system development, **empowering users to communicate with AI systems for their practical demands is critical to unleashing useful AI explanations** (RQ3, Fig 1 C). To this end, we developed two interactive XAI systems that improved the usefulness of AI explanations in terms of human-perceived performance in AI-assisted writing tasks [9] (Fig 1 C1), and model performance in-context learning of large language models [17] (Fig 1 C2), respectively. Overall, we summarize this doctoral research by discussing the limitations and challenges of human-centered useful AI explanations.

2 HUMAN EVALUATION ON XAI USEFULNESS

To investigate if the cutting-edge AI explanation approaches are useful for humans in practice, we first conduct human evaluations on analyzing AI explanations. In particular, we deploy two real-world human studies with MTurk workers on two representative AI explanation methods [10, 13]. First, focusing on image classification tasks, we ask humans to infer the AI model's incorrect labels with a set of post-hoc explanations [10] as shown in Fig 2. We then assess the human accuracy performance to evaluate if AI explanations can be useful for humans to understand the incorrectly predicted labels produced by the image classifiers. The results, collected from 150 online crowd workers with 3800 submissions demonstrate that displaying the AI explanations did not increase, but rather **decreased**, the average guessing accuracy by roughly 10% (Fig 2 4).

On the other hand, in Fig 3, we also target on text classification tasks, where we ask humans to simulate AI model's correct predictions with the help of the model's self-generated inherent explanations (*i.e.,* rationales) [13]. As shown in Fig 3 4, we examine the human simulation accuracy and find that AI explanations that are too short (*e.g.,* with 10% rationale length) **do not help** humans predict labels better than randomly masked texts.

In short, the two human studies on evaluating both post-hoc and inherent explanations show that **AI explanations are not always useful for humans to analyze AI predictions in practice**.

3 DISPARITY BETWEEN AI EXPLANATIONS AND PRACTICAL HUMAN DEMANDS

The aforementioned findings motivate our research in gaining insights into disparities between the status quo of AI explanations and real-world user needs [11]. Specifically, we survey over 200 AI explanation papers[1], summarize the common XAI forms, and compare these forms with the practical user demands represented in the XAI Question Bank [7]. As a result, we observe two dominating findings: *i)* humans request diverse XAI questions across the AI pipeline to gain a global view of AI system (*e.g.,* datasets, models, evaluations, etc), whereas existing XAI approaches commonly display a single AI explanation that merely answers one or several XAI questions. This mismatch can not satisfy diverse XAI

[1]Website of 200+ XAI papers: `https://human-centered-exnlp.github.io/`

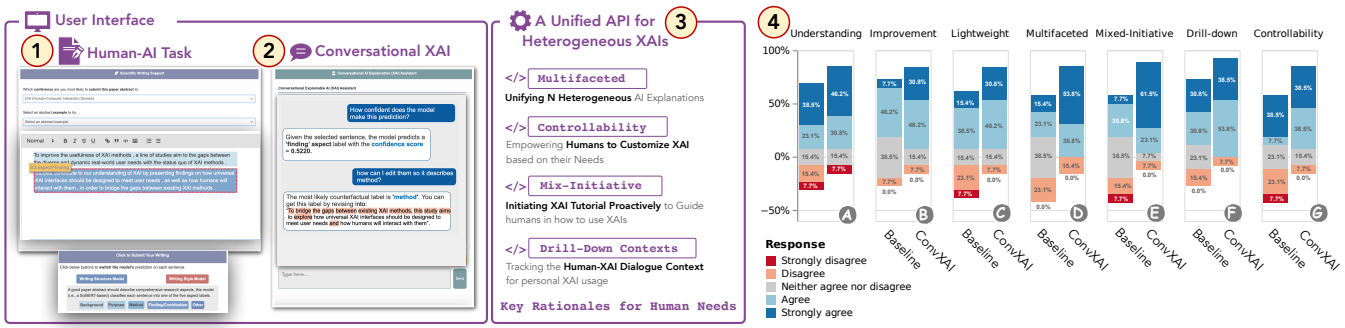
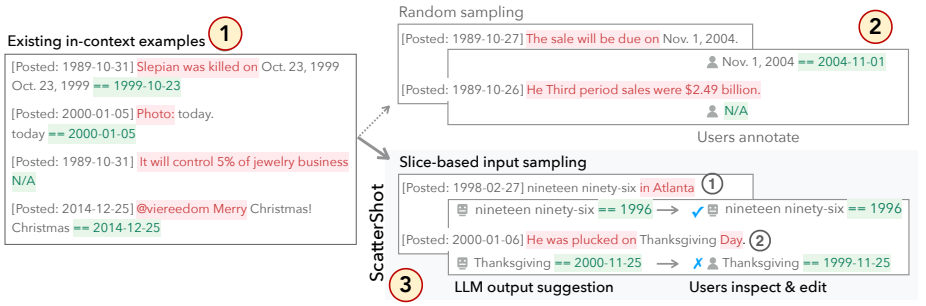


Figure 4: ConvXAI [9] user interface shows the (1) human-AI task and the (2) conversational XAI agent to users for interaction. Particularly, we incorporate (3) four key features for useful XAI collected from formative use studies into ConvXAI. (4) Results show that ConvXAI outperforms the baseline in self-perceived understanding, usefulness, and other metrics.



4

Conditions	ROUGE-L	BLEU-4
Rule-based	78.4	66.7
Random	74.3 ± 3.9	65.4 ± 3.5
ScatterShot	**80.0 ± 3.5***	**69.1 ± 3.1***

Table: Quantitative results comparing ScatterShot with the random baseline on QA-pair, averaged over 10 random seeds. ScatterShot outperformed the baseline on all metrics.
(Student's t-test: *: p<0.05, **:p<0.01).

Figure 5: An overview of how humans use ScatterShot [17] to iteratively collect in-context examples. Given an in-context example set that is likely *underspecifying* the intended functionality (1), ScatterShot extracts typical inputs that are either with novel patterns or difficult in performance (3). Compared with baseline (2), results show (4) ScatterShot helps humans re-allocate annotation budgets towards informative and more explainable examples, increases the in-context function performance.

user needs; *ii)* humans are widely interested in understanding what AI systems can not achieve. Due to the infinite nature of "cannot" inquiries, it is infeasible for XAI developers to specify counterfactual predictions of AI explanations in advance. Consequently, there arises a potential necessity for interactive AI explanations that empower individuals to stipulate the counterfactual predictions to be elucidated.

4 EMPOWER USEFUL AI EXPLANATIONS WITH HUMANS-AI INTERACTIONS

In light of these findings, we argue that empowering users to communicate with AI systems via interactions or even conversations for their practical explanation needs is critical to unleashing useful AI explanations, which is also supported by other literature emphasizing the selective and social nature of explanations [8, 16]. To this end, we explore two research directions, as depicted in Fig 1 C1 and C2, where we develop two interactive AI explanation systems that aim to enable humans to interactively inquire AI explanations for improving human performance [9] and AI model performance [17], respectively, in practical human-AI collaborative systems.

4.1 Conversational AI explanations to improve AI-assisted human writing tasks

On one hand, to bridge the gap between user demands and existing XAI methods, we propose a conversational XAI system, ConvXAI, which incorporates multiple types of AI explanations into a universal XAI dialogue interface and empowers users to ask a variety of XAI questions via the chatbot. More importantly, we identify a set of practical XAI user demands through formative human studies with 7 users of diverse backgrounds, and represent them as four design principles of useful XAI. Specifically, these conversational XAI systems should be able to address various user questions ("multi-faceted"), actively provide XAI tutorials and suggestions ("mix-initiative"), empower users to dig into AI explanations ("context-aware drill-down"), and make flexible customization with details on-demand ("controllability").

Furthermore, we evaluate the ConvXAI system by conducting within-subject user studies with 21 participants. We compare ConvXAI with *SelectXAI–* the conventional GUI-based XAI system that statically displays all the XAIs in a collapsible manner. The findings found that most users perceived ConvXAI to be more useful in **perceived understanding AI writing feedback and improving human writings**. The results also validated the less cognitive load and effectiveness of the four user-oriented design principles. Furthermore, by analyzing the user demands and usage patterns during the tasks, we found that different users prioritize different AI explanations and orders for their needs, their needs change over time, and XAI customization is important for their needs.

Based on the findings, we finally present our speculation into the **core features of useful AI explanations for humans**, which include i) multifaceted XAI to provide diverse explanation perspectives; ii) XAI controllability that enables humans to tailor AI explanations for their customized needs; and iii) proactive XAI tutorial that teaches humans *how to use AI explanations*.

4.2 Interactive AI explanations to improve in-context learning for language models

On the other hand, to improve the AI model performance with example-based explanations (*i.e.,* data examples), we present ScatterShot, an interactive system for building high-quality demonstration sets for in-context learning of large language models. In a nutshell, ScatterShot helps users find informative input examples in the unlabeled data, annotate them efficiently with the help of the current version of the learned in-context function, and estimate the quality of said function. In each iteration, ScatterShot automatically slices the unlabeled data into clusters based on task-specific *key phrases*. Users are then presented with examples of underexplored clusters, or hard examples of explored clusters. Users can either accept correct predictions from the current function or make edits to fix wrong predictions. These additional labels are used to update the in-context function to improve the model performance.

We evaluate ScatterShot both in terms of sampling efficiency and support for human annotators. In simulation experiments, we compare the sampling strategy in ScatterShot to random sampling on two text transformation tasks contemplated in prior work: the data wrangling task illustrated, and rewriting question-answer pairs into logically equivalent pairs in order to evaluate model consistency in Figure 5. In both cases, we find ScatterShot improves performance on corresponding metrics (*e.g.,* Rouge-L, F1) by 4-5 percentage points, with less variance for various values of k demonstrations. Further, we conduct a within-subject user study in which 10 participants build in-context functions for the QA-pair rewriting task either (1) manually, (2) with the ScatterShot interface but a random sampling, or (3) with the fully-featured ScatterShot. We show that ScatterShot's interface alone is an improvement, by offloading input selection and providing sample outputs. Moreover, the sampling strategy in the fully-featured ScatterShot helps users notice diverse input patterns, leading to improvements in the resulting in-context function.

5 LIMITATION AND CHALLENGES

We are excited to contribute to and observe the progress in this field, but acknowledge the long path ahead to achieve useful AI explanations for humans in diverse real-world scenarios. A key challenge is the lack of well-studied benchmarks to define and measure the objective and subjective usefulness of AI explanations across complex use cases. These benchmarks should assess AI explanation impact on human performance (accuracy, efficiency, understanding) and model performance (accuracy, fairness). Interactive AI explanations can be a starting point for future research in expanding the scope of practical AI explanations.

We also hope to inspire more researchers to explore various approaches for making AI explanations useful in practice, such as *building interactive AI explanations to fulfill practical human-AI use cases*, *developing benchmarks to evaluate XAI usefulness*, etc. These efforts will lead to better AI explanations and foster more advanced, fair, trustworthy, and safe human-AI collaborations in general.

6 ACKNOWLEDGEMENT

I would like to express my gratitude to Dr. Ting-Hao 'Kenneth' Huang and Dr. Sherry Tongshuang Wu. I am truly appreciative of their invaluable advice and unwavering support throughout the journey of my doctoral research. Additionally, I express sincere gratitude to all my collaborators who have contributed to the projects I have undertaken, as their constructive insights and expertise have greatly enriched my learning experience.

REFERENCES

[1] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Túlio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *CHI* (2021).

[2] Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Online.

[3] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4443–4458. https://doi.org/10.18653/v1/2020.acl-main.408

[4] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *ArXiv preprint* abs/1702.08608 (2017).

[5] Raymond Fok and Daniel S Weld. 2023. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. *arXiv preprint arXiv:2305.07722* (2023).

[6] Bernease Herman. 2017. The Promise and Peril of Human Evaluation for Model Interpretability. *ArXiv preprint* abs/1711.07414 (2017).

[7] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[8] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[9] Hua Shen, Chieh-Yang Huang, Tongshuang Wu, and Ting-Hao'Kenneth' Huang. 2023. ConvXAI: Delivering Heterogeneous AI Explanations via Conversations to Support Human-AI Scientific Writing. *The 26th ACM Conference On Computer-Supported Cooperative Work And Social Computing - Demonstrations (CSCW Demo)* (2023).

[10] Hua Shen and Ting-Hao Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (AAAI HCOMP)*, Vol. 8. 168–172.

[11] Hua Shen and Ting-Hao'Kenneth' Huang. 2021. Explaining the Road Not Taken. *ACM CHI 2021 Workshop on Human-Centered Explainable AI* (2021).

[12] Hua Shen and Tongshuang Wu. 2023. Parachute: Evaluating interactive human-lm co-writing systems. *ACM CHI 2023 Workshop on In2Writing* (2023).

[13] Hua Shen, Tongshuang Wu, Wenbo Guo, and Ting-Hao Huang. 2022. Are Shortest Rationales the Best Explanations for Human Understanding?. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics.

[14] Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. 2022. Improving fairness in speaker verification via Group-adapted Fusion Network. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7077–7081.

[15] Hua Shen, Vicky Zayats, Johann C Rocholl, Daniel D Walker, and Dirk Padfield. 2023. MultiTurnCleanup: A Benchmark for Multi-Turn Spoken Conversational Transcript Cleanup. *arXiv preprint arXiv:2305.12029* (2023).

[16] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[17] Sherry Wu, Hua Shen, Daniel S Weld, Jeffrey Heer, and Marco Tulio Ribeiro. 2023. ScatterShot: Interactive In-context Example Curation for Text Transformation. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI)*. 353–367.

[18] Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. 2020. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.