Research Statement: From Theory to Practice in Explainable Artificial Intelligence

Shichang Zhang
Department of Computer Science, University of California, Los Angeles

Artificial Intelligence (AI) has pervaded all aspects of our world, from recommending suitable products to customers to assisting doctors in medical decision-making and even helping scientists make discoveries like new drugs and materials. While efforts are constant to enhance AI capabilities, a persistent challenge is the black-box nature of modern neural-network-based AI systems. This obscurity makes it difficult for humans to understand AI, work with AI, and align AI decisions with our values, especially in critical science, business, and healthcare domains, where decisions bear immense significance.

Research goal. The growing integration of AI in critical areas demands a concurrent advancement in Explainable Artificial Intelligence (XAI). I emphasize the paramount importance of XAI. My vision for XAI is seamless communication between humans and AI systems. Future AI should not only be an obscure model making predictions, but a transparent and trustworthy co-worker of humans. However, a gap persists between theoretical advancements in XAI and their real-world applications. Existing XAI methods often cater more to AI specialists than the intended end-users. For instance, technical details like neural network weights or saliency maps of image pixels are often used as explanations, whereas users would benefit more from lucid and domain-specific explanations. A doctor does not need to understand the intricate weights of a deep neural network underlying an AI assistant, but rather how a set of critical vital signs and medications interact to suggest a treatment. The real challenge is translating AI's technical intricacies into domain-specific, user-friendly explanations. In other words, making the explanations more explainable and contextually meaningful.

Reseach experience. As building blocks towards my vision for XAI, I have devised pragmatic XAI models for various high-stake domains in science, business, and healthcare, such as XAI for predicting molecule properties [1], predicting material energy barrier [2], recommendation systems [3], and predicting ICU length-of-stay [4]. For these projects, integrating domain knowledge is vital for bridging the theory-practice gap of XAI. I believe the two important factors for pushing XAI forward to seamless human-AI communication are getting more domain experts involved and developing more powerful AI tools. The powerful tools I have mastered while conducting my research include Graph Neural Networks (GNNs) [5], which can capture complex feature interactions that benefit model explanation, and Large Language Models (LLMs) [6], which provide an easy interface for human-model communication.

The field of cheminformatics has been using AI to predict molecule properties to help new drug discovery, where atoms form the primary input features. However, it is how atoms bond, presenting their interactions in a natural graph structure, encapsulates crucial information regarding molecule properties. My work [1] explains AI models on molecules with a focus on the structure-awareness of the generated explanations. I connect XAI on graphs to game theory by mapping atoms and bonds to game players and their interaction rules respectively. Through computing the Hamiache-Navarro (HN) value [7], my method can extract meaningful subgraph explanations corresponding to important functional groups for predicting molecule properties, with higher fidelity than other XAI methods like the Shapley-value-based approaches.

In material science, AI models like GNNs are also becoming popular, particularly in studying metallic

materials [8]. While these models make predictions on material properties with atomic structure inputs, demystifying the intricate role of atomic structures remains a challenge because of AI obscurity. In my work [2], I introduce a novel GNN to better capture the invariance of atomic structure and explain the energy barrier predictions using gradient analysis and important edge selection. I then work with material scientists to discover the distribution patterns of distances and angles of the selected edges across material samples. My method uncovers the correlation between the energy landscape and atomic structures of metallic materials, consequently validating hypotheses postulated by material scientists and paving the way for better material property understanding and new material discovery.

For business AI models, ensuring simplicity and user-friendliness are even more critical because of the diverse backgrounds of customers. In my collaboration with Amazon [3], I developed an explanatory algorithm tailored for online recommender systems. My solution centers on designing efficient path-finding algorithms on a graph depicting user-item interactions. The derived path explanations offer concise and informative routes on the user-item graph and can be articulately conveyed to customers via natural language. This innovation has been demonstrated to enhance the recommendation's persuasiveness among a user survey, and it is now undergoing integration into Amazon's product recommendation pipeline.

For healthcare problems, the stakes are even elevated, and the predictions and explanations of AI models must be genuinely useful for doctors. In my work [4], I employ XAI techniques to predict and explain the ICU length of stay. My approach is to leverage graph structure learning and attention scores to analyze the patient diagnosis, vital signs, and medications, with a focus on the interaction among these factors. This enables the AI system to assist doctors in discerning the salient factors that influence a patient's duration in the ICU, as well as the complex factor interactions that lead to these suggestions.

Future Research Plans. My overarching ambition is to achieve seamless communication between humans and AIs, to make AIs not merely as a tool, but as a transparent and trustworthy co-worker aligned with human values. This endeavor requires closing the theory-practice gap of XAI by adopting a user-centric approach, offering more contextually meaningful and user-friendly explanations.

Along my journey to this goal, several challenges emerged. Firstly, the essence of effective communication lies in personalization. Just as great teachers tailor explanations based on their students, explanations should be personalized to achieve the desired seamlessness. Yet, the scale and diversity of AI applications and users make this a daunting task. Secondly, related to the first challenge, evaluating the efficacy of these explanations is pivotal. While the widely used evaluations are often metrics centered on AI researchers, the true measure of success should be user-centric and personalized. Large-scale, diverse human evaluations remain scant in XAI. Lastly, deciphering the true causal relationships that underpin the correlations remains a formidable challenge. While AI models are proficient at determining correlations, unveiling the real causal interplay behind decision-making is more profound and more persuasive to end users, especially in domains like healthcare.

Regarding these challenges, the recent advancement of LLMs offers the potential to tackle the first challenge of personalized explanations, as they provide an easy interface for human-model communication, and personalized information can be injected via in-context learning. LLMs can potentially also be used for the second challenge of large-scale explanation evaluation given their human-level language understanding ability in many cases. However, for the third challenge, I think we are yet to reach the point of causal explanation, even with LLMs, as the hallucination problem and the reasoning ability of LLMs are challenging XAI problems as well, which I believe are pre-requisites for causal reasoning. One of my recent explorations is to benchmark LLM reasoning ability with college-level scientific problems [6].

These challenges puzzle me but also fuel my passion and drive for innovation. Addressing them requires a combined effort from cross-disciplinary researchers. I am committed to creating a collaborative space that brings together AI researchers, scientists, industry experts, doctors, ethicists, and more. I aim to transition XAI from an academic endeavor to a broader initiative with universal benefits.

References

[1] **Shichang Zhang**, Yozen Liu, Neil Shah, and Yizhou Sun. Gstarx: Explaining graph neural networks with structure-aware cooperative games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[2] Haoyu Li*, **Shichang Zhang***, Longwen Tang, Mathieu Bauchy, and Yizhou Sun. Predicting and interpreting energy barriers of metallic glasses with graph neural networks. *NeurIPS Workshop on AI for Accelerated Materials Design (NeurIPS AI4Mat)*, 2023.

[3] **Shichang Zhang**, Jiani Zhang, Xiang Song, Soji Adeshina, Da Zheng, Christos Faloutsos, and Yizhou Sun. Page-link: Path-based graph neural network explanation for heterogeneous link prediction. In *Proceedings of the Web Conference (WWW)*, 2023.

[4] Tianjian Guo, **Shichang Zhang**, Indranil Bardhan, and Ying Ding. Predicting icu length of stay: A graph learning-based explainable ai approach. *Workshop on Information Technologies and Systems (WITS)*, 2023.

[5] **Shichang Zhang**, Yozen Liu, Yizhou Sun, and Neil Shah. Graph-less neural networks: Teaching old MLPs new tricks via distillation. In *International Conference on Learning Representations (ICLR)*, 2022.

[6] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, **Shichang Zhang**, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *NeurIPS Workshop on Mathematical Reasoning and AI (NeurIPS Math-AI)*, 2023.

[7] Gérard Hamiache and Florian Navarro. Associated consistency, value and graphs. *International Journal of Game Theory*, 49(1):227–249, 2020.

[8] V. Bapst, T. Keck, A. Grabska-Barwińska, C. Donner, E. D. Cubuk, S. S. Schoenholz, A. Obika, A. W. R. Nelson, T. Back, D. Hassabis, and P. Kohli. Unveiling the predictive power of static structure in glassy systems. *Nature Physics*, 16(4):448–454, April 2020.