Scientific Review for Explainable AI

The overall objective of the research is to develop explainable AI systems to advance machine learning and human-computer interaction. After a review of the proposed approach, scientific merits regarding the research are summarized below.

Scientific Basis and Rationale.

A considerable number of studies have shown that in the use of decision aids and recommender systems of various types, the provision of explanations significantly contributes to user acceptance of and reliance upon computer-provide guidance (e.g., Herlocker, et al., 2000; Tullio, et al, 2007). Furthermore, users like to be provided with explanations, even if they are justifications of system operations expressed in somewhat formal terms (Biran and Cotton, 2017).

The proposed research has the following major innovations, which are potentially transformative and paradigm shifting for machine learning, computer vision, autonomy, HCI and AI: (1) Developing a novel learning regime to achieve both performance and interpretability. (2) Developing a common representation bridging logic deduction and deep learning. (3) Formulating and learning X-AOG, i.e., the dialogue grammar in the explanations.

Competency of Personnel and Adequacy of Proposed Resources.

PI Song-Chun Zhu has extensive experience of managing big projects by serving as PI of two consecutive MURI projects on Scene Understanding [2010-2016] and Commonsense Reasoning [2015-2020], and MSEE project [2010-2015] on visual analytics and SIMPLEX [2015-2018] on robot autonomy. He serves as general chair for CVPR2012 & CVPR2019, and on many editorial board and panels.

PI Zhu's lab has been developing explainable AI systems since 2017. Members on the team have long track record of successful and effective collaborations. The team will have regular weekly meetings via video-conference to coordinate progress. A common Github site and dashboard will be used to document and manage the project progress on-line, sharing code and data, as we have been doing currently. Software will be modular and open source, version control will be used to assigning tasks to various members of the team.

Appropriateness of the Proposed Study Design.

The team would conduct both crowd-sourcing and in-lab experiments to evaluate the effectiveness of the explanation interface and the dialogue. Different evaluation metrics would be applied based on evaluation criteria including correctness in recognizing the type of intents of users' explanation requests, relevancy of explanations, fidelity to the interpretable model's behaviors, completeness of the explanation and efficiency and effectiveness of the explanation dialogue. Standard statistical tests (e.g., ANOVA) will be used to measure the performance of the XAI system, based on the above-mentioned criteria.

Reporting of unanticipated problems involving risks to subjects or others (UPIRTSOs), suspensions, terminations, serious adverse events (SAE), and serious or continuing non-compliance would be made in accordance with the Department of Defense (DoD) and UCLA HRPP regulations.

Reference.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241–250). ACM.

Tullio, J., Dey, A. K., Chalecki, J., & Fogarty, J. (2007). How it works: a field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 31–40). ACM.

Biran, O., & Cotton, C. (2017). Explanation and Justification in Machine Learning: A Survey. *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI).*

Hongquan Xu
Professor and Chair
UCLA Department of Statistics

Signature: Hongquan Xu
Date: 11/1/19