Effects of AI Explanations on Users with Different Levels of AI Knowledge

Akihiro Maehigashi
Shizuoka University, Shizuoka, Japan

Seiji Yamada
Kanagawa University, Kanagawa, Japan

Abstract

This study investigated how different types of AI explanations, information provided by AI systems to support users' interpretation of AI behavior, influence interpretability and trust in AI among users with different levels of AI knowledge. We compared three explanation types that varied in their degree of grounding in the underlying model behavior: convolutional neural network (CNN) label-based explanations, Grad-CAM heatmap explanations, and large language model (LLM)-generated textual explanations. The results showed that users with lower AI knowledge perceived LLM-generated explanations as more interpretable and trustworthy than CNN and Grad-CAM explanations. At the same time, trust in CNN and Grad-CAM explanations remained relatively high despite limited understanding. Notably, LLM-generated explanations also increased interpretability and trust among users with higher AI knowledge, despite being only weakly grounded in the underlying model behavior. These findings highlight the importance of designing AI explanations that meaningfully connect model behavior, explanatory representations, and user interpretation.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*