Abstract

The integration of Artificial Intelligence (AI) in medical decision-making presents unprecedented opportunities for enhancing diagnostic accuracy and decision support. However, their opaque 'black-box' nature creates significant challenges, particularly in fostering trust among users such as medical experts and patients. Explainability is a crucial factor in addressing these challenges, yet its effects remain ambiguous. This dissertation critically examines the beneficial and adverse impacts of AI explanations in medical contexts, providing a nuanced perspective on their role in AI-advised medical decision-making.

The dissertation comprises one conceptual and three empirical studies investigating how explanations influence medical experts and patients. Empirical findings reveal that explanations enhance perceived transparency, usefulness, and causal understanding, fostering greater adoption and trust in AI systems. They also serve as a 'vaccine' against discontinuance, helping users maintain confidence even when AI errors occur. Furthermore, explanations alleviate privacy concerns for some users by reducing uncertainty about data use. However, the dissertation also uncovers significant adverse effects. Explanations can inadvertently increase cognitive load, overwhelm users in high-stakes environments, and may amplify privacy concerns by drawing attention to sensitive data processing. These findings underscore the double-edged nature of explanations and the necessity for careful implementation to maximize benefits while minimizing risks.

To bridge theory and practice, this dissertation develops a procedural model that guides the integration of explanations across the pre-use, use, and post-use phases of user interaction with explanations. The model demonstrates its utility in hypothetical scenarios, such as clinical decision support systems for radiologists and symptom checker applications for patients, offering tailored strategies for balancing transparency, cognitive load, and privacy considerations.

Contributing to the Human-AI interaction literature and advancing discussions on the dark sides of information systems, this research highlights the strategic importance of explanations in fostering long-term adoption of AI technologies. It provides theoretical insights, empirical evidence, and actionable recommendations for designing explainable AI systems that align with user needs and regulatory requirements. By emphasizing both the benefits and potential pitfalls of AI explanations, this dissertation advances a balanced, user-centered approach to developing trustworthy and effective AI systems in medical applications.