Explainability Shapes How People Respond to Errors in AI and Human Decision Support and Buffers Future Trust

Zuming Zhang

University of Sydney, Sydney, Australia

Bruce D. Burns

University of Sydney, Sydney, Australia

Abstract

The current study examines the effects of advice explainability on advice judgment and post-error response when advice is provided by a goal-specific AI system versus a human expert in high-stakes decision-making contexts. Across four scenarios, participants were presented with advice that was either accompanied by justification or not, and the advice was subsequently revealed to be incorrect. Results showed that prior to the error revelation, higher explainability did not affect perceived advice correctness, perceived advice understandability, or willingness to follow the advice, nor did it eliminate the preference for human advisors over AI systems. Instead, explainability most clearly shaped post-error responses. Providing justifications reduced responsibility attributed to the advisor, increased perceived error understandability, and lowered unwillingness to follow future advice. This pattern suggests that justifications may invite excuse generation and external attributions, making failures appear more understandable, reducing blame toward the advisor, and thereby buffering future trust.

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