Contribution ID: **646** Type: **Mini-talks**

Cognitive Biases in Human-AI Interaction: Psychological Implications for Explainable Artificial Intelligence

Friday, September 12, 2025 5:50 PM (10 minutes)

As Artificial Intelligence (AI) becomes increasingly embedded in high-stakes domains such as healthcare, law, and public administration, automation bias (AB) -the tendency to over-rely on automated recommendations- has emerged as a critical challenge in human-AI collaboration. While previous reviews have examined AB in traditional computer-assisted decision-making, research on its implications in modern AI-driven work environments remains limited. To address this gap, this research systematically investigates how AB manifests in these settings and the cognitive mechanisms that influence it.

Following PRISMA 2020 guidelines, we reviewed 17 peer-reviewed studies (2014–2024) from SCOPUS, ScienceDirect, PubMed, and Google Scholar. Specifically, while Explainable AI (XAI) and transparency mechanisms are designed to mitigate AB, evidence suggests they may inadvertently reinforce it by fostering misplaced trust, particularly among users with lower AI literacy. Additionally, individual differences such as professional expertise, cognitive engagement, and prior exposure to AI significantly shape susceptibility to AB.

Traditional perspectives attribute AB to over-trust in automation, assuming users perceive AI-generated outputs as inherently reliable. However, our critical review presents a more nuanced view. While confirming some prior findings, it also sheds light on new aspects, reframing AB as a bias shaped by both cognitive and attitudinal factors, as well as task characteristics. AI literacy, professional expertise, cognitive profile, trust dynamics, task verification demands, and explanation complexity all play a role.

Finally, we propose explanation design strategies that actively promote critical engagement and independent verification, offering both theoretical and practical contributions to bias-aware AI development.

If you're submitting a symposium talk, what's the symposium title?

If you're submitting a symposium, or a talk that is part of a symposium, is this a junior symposium?

Primary authors: CONTI, Daniela (Università di Catania); Mr ROMEO, Giuseppe (Università di Catania)

Presenter: CONTI, Daniela (Università di Catania)

Session Classification: Reasoning and abstract cognition

Track Classification: Reasoning and abstract cognition