Mitigating Automation Bias in Generative AI Through Nudges: A Cognitive Reflection Test Study

Wingerter et al. (2025)

Context

Automation bias refers to over-reliance on GenAI-outputs, accepting them without scrutiny. This experiment (n = 190) used the Cognitive Reflection Test (CRT) to test whether faulty AI answers trigger this bias, and if a simple warning nudge can counteract it. When AI gave faulty answers, users were biased and got fewer than half the questions right compared to those not using AI. A simple warning nudge almost doubled performance relatively, but did not push it above the no-AI control group. Crucially, AI literacy alone didn't have a significant impact, suggesting organisations cannot rely on employee training without embedding safeguards into the AI tools themselves.

Key Insights

Incorrect AI outputs often override users' own reasoning

- Participants receiving incorrect AI suggestions answered fewer than half as many questions correctly as those working without AI, reducing performance to approximately 39% of the control group
- Users defaulted to accepting AI outputs rather than engaging in deliberative reasoning, consistent with dual-process theory (System 1 overriding System 2)

Even AI literacy does not protect against automation bias

- Participants who reported higher AI knowledge and experience performed no better at spotting faulty AI outputs than those with lower AI literacy
- Possible explanations include confirmation bias, anchoring on the AI's initial output and overconfidence/ the Dunning-Kruger effect (when users with limited knowledge overestimate their skills in understanding or prompting the AI due to the confident and fluent outputs the AI generates)
- This implies that even user familiarity with AI does not build the critical scrutiny needed to consistently catch errors without structural assistance for just-in-time reflection

A simple warning nudge nearly doubled performance, but only back to the baseline

- A prominently displayed warning reminding users that AI outputs may contain errors almost doubled accuracy compared to the faulty AI condition (+87%)
- However, the nudge only restored performance to the level of the no-AI control group, not above it
- This suggests nudges can counteract automation bias but are insufficient on their own to improve reasoning quality beyond what people achieve unaided

Implications

- How do you maintain oversight you're not just building employee's AI literacy but that it translates into continous critical AI output evaluation in their day-to-day work? Are tools in place for this?
- Where could simple interface-level nudges (warnings, verification prompts) be built into AI-usage?
- If even such nudges only restore performance but don't improve it, what could an additional structured review step look like before your employees accept AI-generated recommendations?

Wingerter, T. L., Straub, T., & Schweitzer, S. (2025). Mitigating automation bias in generative AI through nudges: A cognitive reflection test study. Procedia Computer Science, 270, 2106-2114. https://doi.org/10.1016/j.procs.2025.09.331