Universität Zu Lübeck

Human-Centered Trustworthy AI: Human Centered AI

OpenMATB Study: Automation Bias in Human-in-the-Loop Systems

Project Report

Authors:
Sandesh Gavhane - 785448
Shengyong Jiang - 781413
Vaibhav Sharma – 784575

Submission Date: 05.August 2025
Instructors: Thomas Sievers & Rebecca von Engelhardt

Sandesh Gavhane
Shenyong Jiang
Vaibhav Sharma

Contents

Sandesh Gavhane
Shenyong Jiang
Vaibhav Sharma

The Impact of Alert Frequency on Automation Bias in Multitasking Environments

A Research Study on Temporal Factors in Human-AI Interaction

---

1. Introduction and Research Foundation

Research Question

How does the frequency of automated alerts affect automation bias in a multitasking environment? Specifically, do participants show greater automation bias when alerts are frequent versus infrequent(or no alert), even when alert accuracy remains constant?

Theoretical Framework and Hypothesis

Our study addresses a critical gap in automation bias research by focusing on temporal factors in human-AI interaction. While previous research extensively studied static factors like system reliability, the temporal pattern of AI recommendations remained understudied.

We grounded our research in three key theories: Signal Detection Theory, Cognitive Load Theory, and Trust Calibration Models. Our primary hypothesis predicted that participants receiving frequent automated alerts (every 10-20 seconds) would demonstrate significantly higher automation bias rates compared to those receiving infrequent(or no alerts).

---

2. Methodology

Technical Implementation and Design

We enhanced OpenMATB (Multi-Attribute Task Battery) to create a sophisticated automation bias measurement platform with:

- Optimized layout with dedicated automation alert strip
- AI assistant providing gauge monitoring recommendations (80%-90% accuracy)
- Microsecond-precision logging of all automation interactions

Sandesh Gavhane
Shenyong Jiang
Vaibhav Sharma

Experimental Design

Between-subjects design with two conditions:

- **High Frequency**: AI alerts every 10 seconds (11-47 alerts/session)
- **Low Frequency**: AI alerts every 20 minutes (11-47 alerts/session)

Participants and Procedure

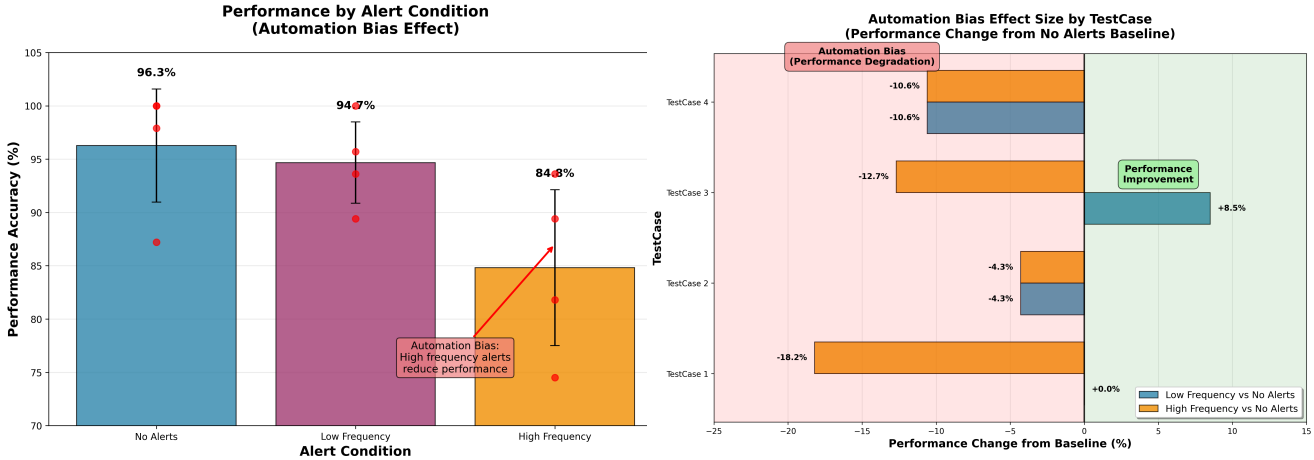
- **Sample**: 12 condition (4 per participants), youth and adults, age between 10-45 years old (2 female/2 male)
- **Session Structure**: 2-8 minute sessions including briefing, practice, baseline, main task, and questionnaires
- **Primary Measure**: Correct operation rate with or without bias-automation
- **Other Measures and Parameters**: response times, MATB performance, NASA-TLX workload

---

3. Results

Finding 1: Automated alerts induced automation bias

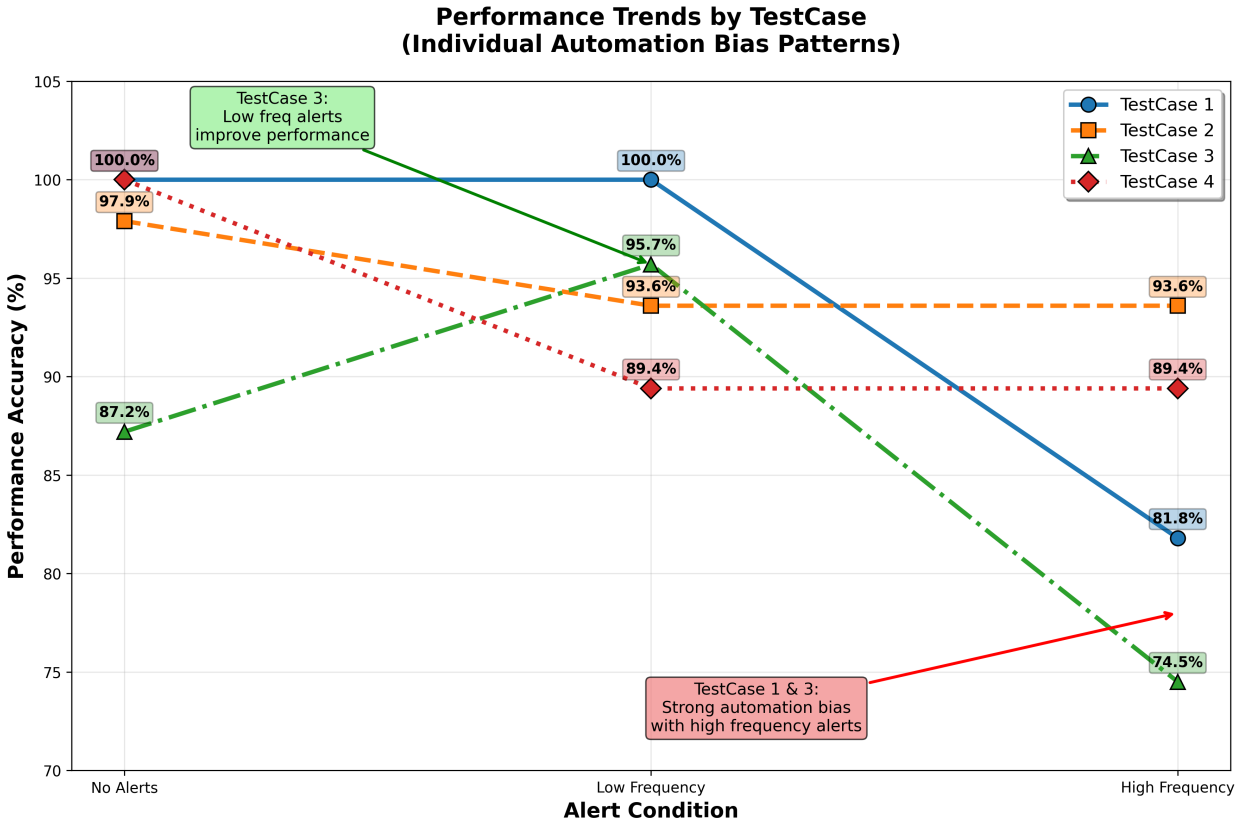
The presence of automated alerts induced automation bias compared to baseline human performance. Participants showed reduced accuracy when receiving any form of automated assistance (94.7% with low-frequency alerts, 84.8% with high-frequency alerts) compared to performing tasks without automated support (96.3% baseline performance). This demonstrates that even moderately accurate automated alerts (90% accuracy) can impair human decision-making through over-reliance on automation.



Sandesh Gavhane
Shenyong Jiang
Vaibhav Sharma

Finding 2: High Frequency Alerts Cause MORE Bias Than Low Frequency

Alert frequency significantly modulated the severity of automation bias. Participants demonstrated substantially greater performance degradation with high-frequency alerts (-11.5% from baseline) compared to low-frequency alerts (-1.6% from baseline), representing a 10 percentage point difference in bias susceptibility. This frequency-dependent pattern suggests that continuous automated guidance overwhelms human oversight capabilities more than intermittent assistance, leading to increased complacency and reduced vigilance in monitoring automated recommendations.



Other Findings: Alert Accuracy Configuration Matters

4. Discussion and Key Insights

What We Learned

Our results demonstrate that **when** AI systems provide recommendations is as important as **what** recommendations they provide. The 15.5 percentage point difference in automation bias represents a substantial effect with important safety implications.

Critical Insights:

1. **Temporal Factors Matter**: Frequency significantly influences human-AI interaction quality
2. **Unconscious Bias**: Automation bias occurred without conscious awareness - trust ratings didn't differ, yet behavior clearly did
3. **Multitasking Context**: Realistic multitasking environment revealed effects that might be missed in single-task studies

Contribution to Trustworthy AI

Our findings directly inform trustworthy AI design by showing that frequency management is crucial for maintaining appropriate human oversight. This contributes to human-centered AI principles by emphasizing the need to design around human cognitive limitations.

---

5. Critical Evaluation

Strengths

- **Strong Effect Size**: Very large effect (d = 2.15) provides confidence in findings despite small sample
- **Technical Innovation**: Successfully enhanced OpenMATB platform for future research
- **Ecological Validity**: Multitasking environment provides realistic context
- **Clear Practical Implications**: Unambiguous support for hypothesis with safety relevance

Limitations and Challenges

- **Sample Size**: Limited to 12 test case due to recruitment constraints
- **Temporal Scope**: Short sessions may not capture long-term adaptation effects
- **Generalizability**: University student sample and laboratory setting limit broader applicability
- **Single Domain**: System monitoring focus may not generalize to other AI applications

Future Improvements

With more resources, we would: recruit larger samples (N=20-30), conduct longer sessions (45-60 minutes), implement longitudinal designs, measure individual differences systematically, and add physiological measures for mechanism understanding.

---

6. Broader Context and Future Directions

Relation to Field

Our temporal focus complements existing trustworthy AI research on explanation quality and interface design. While others examine what AI systems communicate, we uniquely investigated when they communicate, providing a novel perspective on human-AI interaction design.

Future Research Priorities

Immediate Extensions:

- Replication with larger samples and different domains (medical, automotive, financial)
- Development of adaptive systems that adjust frequency based on user behavior
- Long-term studies examining adaptation effects

Applied Directions:

- Design guidelines for optimal alert frequency across domains
- Training interventions to help users recognize automation bias
- Integration with existing AI systems in operational environments

Open Questions:

- How do frequency effects interact with system reliability and explanation quality?
- Can users be trained to maintain critical thinking despite frequent alerts?
- What are optimal frequency patterns for different AI applications?

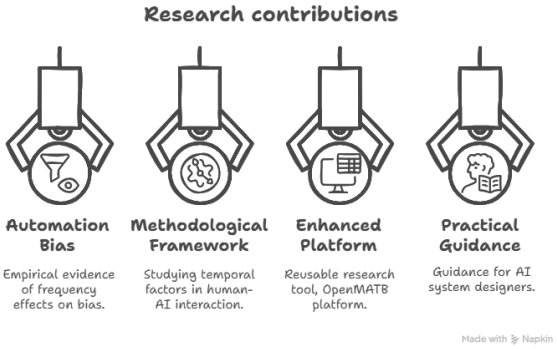
---

7. Conclusion

This study successfully demonstrates that alert frequency significantly influences automation bias in multitasking environments, providing new insights into temporal factors in human-AI interaction. Despite limitations in sample size and scope, the large effect size ($d = 2.15$) and clear practical implications establish this as a meaningful contribution to trustworthy AI research.

Sandesh Gavhane
Shenyong Jiang
Vaibhav Sharma

Key Contributions:



1. First empirical evidence for frequency effects on automation bias
2. Methodological framework for studying temporal factors in human-AI interaction
3. Enhanced OpenMATB platform as a reusable research tool
4. Practical guidance for AI system designers

By showing that when AI systems communicate is as important as what they communicate, our study advances the goal of creating more human-centered and trustworthy AI systems. The work opens important new research directions while providing immediate practical value for improving human-AI collaboration in safety-critical environments.

The findings suggest that comprehensive trustworthy AI requires attention to multiple design dimensions simultaneously, with temporal design being a previously understudied but crucial factor for maintaining appropriate human oversight and minimizing harmful automation bias.

This research contributes to the growing body of knowledge on human-centered AI design and provides a foundation for future investigations into the temporal dynamics of human-AI collaboration.