CSET CENTER for SECURITY and EMERGING TECHNOLOGY

Summary of “AI Safety and Automation Bias”

Automation bias is the tendency for an individual to over-rely on an automated system. It can endanger the successful use of AI by eroding the user’s ability to meaningfully control an AI system. Automation bias can lead to accidents, errors, and other adverse outcomes when individuals and organizations favor the output or suggestion of the system, even in the face of contradictory information.

This study provides a three-tiered framework to understand automation bias at the level of the user, within technical designs, and organizational processes. It presents case studies to illuminate lessons learned and recommendations:

User Bias, Tesla Case Study
Factors influencing bias: ● User’s personal knowledge, experience and familiarity. ● User’s degree of trust and confidence in themselves and the system.
Lessons learned from case study: ● Disparities between user perceptions and system capabilities contribute to bias and may lead to harm.
Recommendation: ● **Create and maintain qualification standards for user understanding.** User misunderstanding of a system’s capabilities or limitations is a significant contributor to incidents of harm.

Technical Design Bias, Airbus vs. Boeing design philosophies Case Study
Factors influencing bias: ● The system’s overall design and user interface.
Lessons learned from case study: ● Even with highly trained, systems interfaces contribute to bias. ● Different design philosophies have different risks. No one approach is perfect but all require clear, consistent communication and application.
Recommendation: ● **Value and enforce consistent design and design philosophies that account for human factors, especially for systems likely to be upgraded.** When necessary, justify and make clear any departures from a design philosophy to legacy users. Where possible, develop common design criteria and consistently communicate them (either through organizational policy or industry standard).

Organizational Policies and Procedure Bias, Army Patriot Missile System vs. Navy AEGIS Combat System Case Study
Factors influencing bias: ● Organizational training, processes, and policies.
Lessons learned from case study: ● Organizations can employ the same technologies differently, based on protocols, operations, doctrine, training, and certification. Choices in each of these areas of governance can embed automation biases. ● Organizational efforts to mitigate automation bias can be successful, but mishaps are still possible, especially when human users are under stress.
Recommendation: ● Where autonomous systems are used by organizations, **design and regularly review organizational policies appropriate for technical capabilities and organizational priorities**. Update policies and processes as technologies change.

Across these three case studies, it is clear that "human-in-the-loop" cannot prevent all accidents or errors. Properly calibrating technical and human fail safes for AI, however, poses the best chance for mitigating the risks of using AI systems.

For more information:

- Download the report: https://cset.georgetown.edu/publication/ai-safety-and-automation-bias/
- Contact us: cset@georgetown.edu