AI Guardrails for Human Use:
Implications for High-Consequence Systems

SA Technologies

- **AI is still not that intelligent**
  - Unable to handle many unforeseen (unlearned) situations
  - Creates difficulties in understanding what it is doing and why, whether it is reliable in current circumstances

  ← Provide Real-time AI Transparency

- **AI has inherent biases**
  - Limitations of training data (often hidden), may not generalize well
  - Simultaneously AI biases attention and decisions of human users (confirmation bias) leading to additional errors

  ← Expose & Avoid AI Bias

- **People cannot effectively overcome AI limits when out-of-the-loop**
  - Situation awareness is lower with automation and AI, degrading take over performance

  ← Provide Safe Fallback State

- **Ultimate success of AI rests on its ability to work effectively with humans in the system**
  - Must evaluate effects on human performance as a key part of system development and testing

  ← Test Joint Human-AI System

For more information: https://www.hfes.org or Contact: mica@satechnologies.com