Institut für Steuerungs- und Regelungssysteme
Professur für Vernetzte Sichere Automatisierungstechnik
Univ.-Prof. Dr.-Ing. Mike Barth
Fritz-Haber-Weg 1 (Geb. 30.33), 76131 Karlsruhe

Approve. Approve. Approve.

Designing Human Oversight for Agentic AI in Industrial Automation

Masterthesis

You design, prototype and evaluate architectures and interfaces to improve applicability of agentic AI to industrial automation by improving their verifiability.

Motivation

Agentic AI systems no longer just answer questions. They plan, call tools, execute code, change configurations and chain dozens of steps together. Asked how we keep them safe, everyone gives the same answer: keep a human in the loop. In practice that means an approval dialog, a diff to confirm, a checkbox before the agent acts.

The uncomfortable question is whether anyone is still looking.

Parasuraman and Manzey showed that operators of reliable automation gradually shift their attention elsewhere and sample the automated output less and less often. Vassilev extends the argument to AI: as agentic systems scale in speed, volume and opacity, human review does not scale with them. The reviewer clicks approve because forty approvals are queued and the previous thirty-nine were fine. What remains is a control that exists on the architecture diagram but no longer does any work, at a time when Article 14 of the EU AI Act demands oversight that people can genuinely exercise. Yet human-in-the-loop is still designed as a UI afterthought: a modal with Allow and Deny. This thesis takes the other route and asks how systems must be built through risk-based interruption, aggregated behavioural summaries, and reversibility and post-hoc audit in place of pre-approving everything. You will design such mechanisms, build them into an agentic prototype, and measure whether real people catch more of the mistakes that matter.

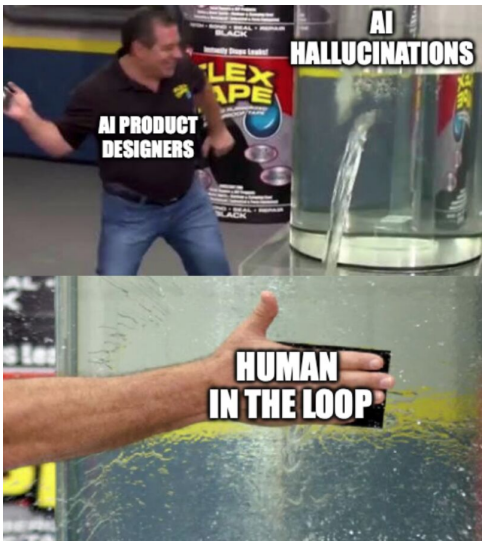

Abbildung 1: Rest of IRS-VSA

Goals

- Survey how human-in-the-loop is realised in current agentic systems and derive a structured overview of oversight points, interaction patterns and their assumptions.
- Design solutions, e.g. a risk-adaptive interruption policy, an attention-aware review interface, or a reference architecture with escalation, sampling and rollback.
- Integrate design into industrial automation agentic setting so that it can be exercised by real users.
- Run study with injected faulty agent actions and measure detection rate, review effort, trust calibration and over- and under-reliance against conventional approval baseline.

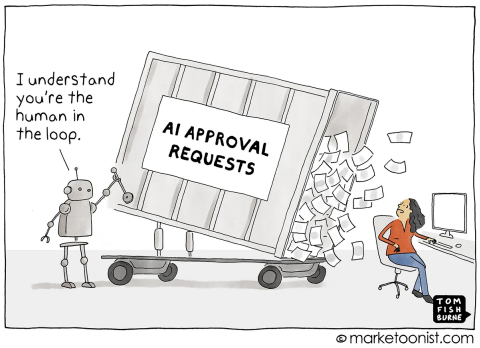

Abbildung 2: Problem Statement

Betreuer
Marwin Madsen, M. Sc.
Build. 30.33, Room 118
Phone: 0721/608-42642
marwin.madsen@kit.edu

Thesis: Masterthesis
Datum der Ausschreibung: 01.09.2026
Tags: *Agentic AI, Security, Human-in-the-Loop, Industrial Automation*