Evaluating Human-AI Interfaces in Logistics: Impact on Cognitive Workload and Usability

Evangelia Fotidou, Lampros Giotis, Christos Katsanos

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

Introduction

- **Background Statement:** Integrating AI algorithms into logistics decision-support systems relieves operators of complex manual calculations (e.g., bin-packing, dynamic rerouting). However, opaque "Black Box" automation often leads to trust calibration failures, inducing either overtrust (automation bias) or undertrust (system disuse).
- **Research goal**: Investigate how an Explainable AI interface designed according to Microsoft HAX guidelines impacts operator cognitive workload (NASA-TLX), usability (SUS, Time on task, User errors), and calibrated trust compared to an unassisted baseline interface.

Methodology (1) → HAX-Grounded Interface Design

Evaluates two interface paradigms implemented within a single-page application framework:

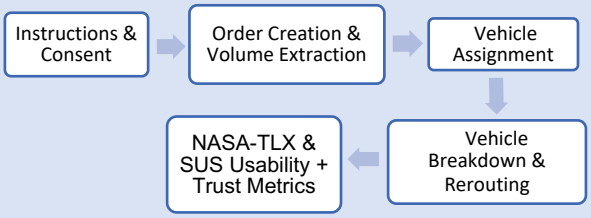
- **Baseline UI:** Unassisted manual fleet filtering, capacity checks, and manual disruption handling.
- **AI-Enhanced UI:** Implementation of Microsoft HAX guidelines:
  - **Initially (G1, G2-A):** Setting realistic system boundaries with explicit prompt cues and displaying approximate ETA ranges instead of deterministic guarantees to avoid false precision.
  - **During Interaction (G3, G4):** Event-driven proactive recommendations and context-aware filtering (visualizing solely vehicles with sufficient cubic capacity).
  - **When Wrong (G7, G11-A):** Local Explanations detailing ETA and volume constraints alongside on-demand override/re-routing controls for Human-in-the-Loop governance.
  - **Over Time (G12):** Temporary state caching to exclude rejected suggestions and prevent repetitive calculation errors.

Participatory Design: 1 logistics professional with >7 years of experience (continuously involved in co-design) and 5 logistics professionals providing iterative feedback across design stages.

Methodology (2) → Evaluated User Tasks

Dispatcher workflow structured into three sequential sub-tasks:

- **Sub-task 1:** Order Intake
- **Sub-task 2:** Capacity Allocation
- **Sub-task 3:** Crisis Recovery

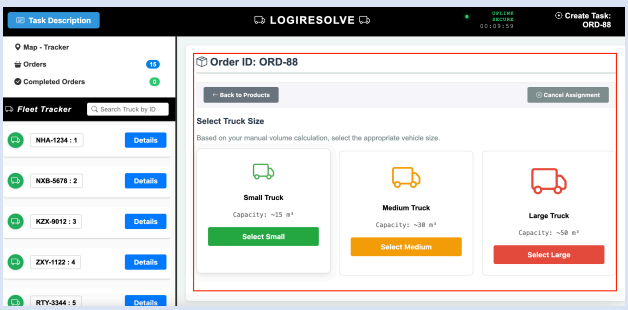


Methodology (3) → Experimental Setup & Study

- **Study Design:** Within-subjects experimental design with counterbalanced order of interface conditions.
- **Participants:** 22 professional logistics operators and fleet coordinators (N = 22, within-subjects counterbalanced design).
- **Evaluation Status:** Completed empirical evaluation.
- **Evaluation Instruments & Metrics:**
  - Cognitive Workload: NASA-TLX across 5 sub-scales (0–100 scale).
  - System Usability: System Usability Scale (SUS) questionnaire (0–100 score).
  - Human-AI Interaction Metrics: Perceived Transparency (G11), Uncertainty Awareness (G2), and Calibrated Trust (5-point Likert scales).

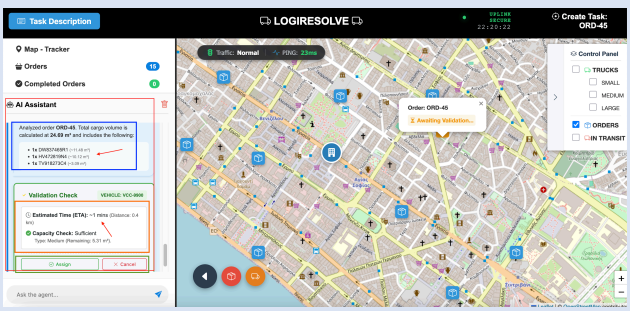
System Interface Showcase

(Baseline UI):



- Unassisted mental calculations
- High extraneous cognitive load & low decision transparency

(AI-Enhanced UI):



- G1: Capability Boundaries
- G2-A: Match Level of Precision & Uncertainty Display
- G11-A: Local Explanations
- G7: Human-in-the-Loop

Hypotheses & Outcomes

Hypotheses	Outcomes
H1: Workload Reduction (NASA-TLX)	**H1: Confirmed** ($p < 0.001, r = 0.765$): 47.2% total workload reduction, ($M = 26.40$ vs. 50.00).
H2a: Usability (SUS) **H2b:** Task Completion Time **H2c:** Error reduction	**H2a: Confirmed** ($p = 0.021, r = 0.348$), $M = 91.36$ ('Best Imaginable') vs. $M = 77.84$ ('Good'). **H2b: Non-significant** ($p = 0.84$): 297.8s vs. 338.1s (speed – accuracy tradeoff). **H2c: Confirmed** ($p = 0.003, r = 0.449$): 72.7% errors reduction ($M = 0.41$ vs. 1.50).
H3: Trust Calibration & XAI Acceptance	**H3: Confirmed.** High transparency ($M_{transparency} = 4.45 / 5$) and confidence ($M_{trust} = 4.27 / 5$).
H4: Automation Bias Mitigation	**H4: Confirmed.** ETA ranges effective ($M = 4.36 / 5$); balanced reliance ($M = 3.23 / 5$).