LUMSA International Research Center for Artificial Intelligence Management

LICAIM Research Project Summary Sheet

1. Title of the Research Project

Measuring trust erosion from AI hallucinations: A comparative survey across critical professional domains

2. Synopsis

This project investigates how artificial intelligence (AI) hallucinations affect professional trust, decision-making attitudes, and perceived reliability across multiple critical domains. While current research on AI trust typically focuses on single sectors, generic attitudes, or conceptual analyses, there is a lack of empirical, cross-professional studies examining how domain-specific hallucinations influence trust erosion in real-world contexts.

The project develops a unified, scenario-based survey instrument designed to assess how professionals react to different forms of AI-generated inaccuracies, including factual errors, fabricated data, and unsupported predictions. The survey is administered across several high-stakes professions that rely on accuracy and accountability (e.g., physicians, teachers, veterinarians, commercial practitioners, scientific professionals). Rather than studying individual professions, the project aims to compare patterns and thresholds of trust degradation across domains.

The expected contribution is twofold. First, it provides empirical evidence on how hallucinations impact trust in AI systems in settings where reliability is crucial. Second, it offers actionable insights for AI management, particularly in human-AI interaction design, risk communication, and responsible deployment strategies aligned with LICAIM's mission. The findings will support the development of frameworks for trustworthy AI implementation in professional environments, informing policy, training, and future interdisciplinary research.

3. Objectives

Main Objective

To measure and compare how AI hallucinations erode trust across different professional communities and to identify common and domain-specific patterns of response.

Specific Objectives

1. To develop a validated survey instrument for assessing trust erosion triggered by AI hallucinations.
2. To collect and analyze data from multiple professional groups using parallel, domain-adapted scenarios.
3. To identify dimensions of trust most sensitive to hallucinations (accuracy, transparency, responsibility, perceived competence).

4. To compare cross-domain differences and similarities in reactions to hallucinations.

4. Theory and Methods

Theoretical Framework

The project draws on the literature of AI trust, human-AI interaction, error typology (particularly hallucinations in LLMs), decision psychology, and sociotechnical systems. Key frameworks include trust calibration, risk perception, and accountability attribution in AI-assisted decision contexts. It also integrates emerging work on professional vulnerability to automation and misinformation generated by AI systems.

Methodology

Research Design: Quantitative, cross-sectional, scenario-based survey; exploratory-explanatory design.

Data Collection:

- Standardized online questionnaire with domain-specific scenarios of AI hallucinations.
- Anonymous responses via secure internal platform.
- Professional groups initially targeted: primary and secondary school teachers, general practitioners (primary care physicians), veterinarians, commercial practitioners, scientific professionals.

Data Analysis: Data analysis will be conducted through a combination of descriptive and inferential statistical techniques. The first stage includes computing descriptive indicators to summarize trust levels, perceived accuracy, responsibility attribution, and reactions to AI hallucinations within each professional group. Reliability analyses will assess the internal consistency of the trust-related constructs. Subsequently, exploratory factor analysis will be employed to identify underlying components of trust erosion, allowing the research team to verify whether the instrument captures distinct dimensions such as accuracy sensitivity, transparency perception, and behavioural intention. Comparative analyses, including ANOVA or multivariate approaches where appropriate, will be used to evaluate differences across professional domains. This analytical strategy is designed to determine both common patterns and domain-specific reactions to AI hallucinations, providing a robust empirical foundation for the interpretation of cross-sector trends.

Sample/Cases: The study aims to collect data from multiple professional groups that operate in high-stakes environments and rely on accurate information for decision-making. The initial dataset consists of thirty secondary school teachers who have already completed the survey. Additional data collection is underway among general practitioners, veterinarians, commercial practitioners, and scientific professionals. Each group is expected to reach a minimum of thirty respondents to ensure comparability and statistical robustness. Participants are recruited through voluntary and anonymous online surveys, with professions selected not as the primary object of study but as application contexts that allow cross-domain comparison. This sampling approach enables the project to explore how trust erosion from AI hallucinations manifests across diverse yet critical sectors of professional practice.

5. Current Status

Completed Activities

- Research design and conceptual framework defined.
- Survey instrument developed and piloted.
- GDPR-compliant data collection infrastructure implemented.
- First dataset collected (30 teachers).
- Adaptation of survey for additional professions completed (physicians, veterinarians, commercial practitioners, biological/scientific professionals).

Ongoing Activities

- Data collection across additional professional sectors.
- Refinement of theoretical framing for LICAIM integration.
- Preparation of dissemination materials for LICAIM and academic venues.

Timeline Progress

Approx. 35% complete, on schedule

Preliminary Findings (if applicable)

Preliminary teacher responses indicate substantial uncertainty regarding transparency, attribution of responsibility, and trust recovery after AI hallucinations. Patterns suggest strong sensitivity to factual errors and predictive inaccuracies.

6. Next Steps

Short-term (next 3 months)

1. Complete data collection for all target professional groups.
2. Conduct reliability and factor analyses.
3. Begin drafting of full research manuscript.

Medium-term (3-6 months)

1. Perform advanced comparative analyses across domains.
2. Derive managerial implications for trustworthy AI in professional practice.
3. Prepare conference abstract and internal LICAIM seminar presentation.

Long-term (6-12 months)

1. Submit manuscript to an international journal (e.g., *AI* - MDPI, or an equivalent high-impact interdisciplinary outlet).
2. Develop policy-oriented recommendations for AI trust management.
3. Produce a LICAIM project report for dissemination on the Center's website.

Expected Outputs

- Journal article
- Conference presentation

Resource Needs (if any)

None at present; project currently sustainable with existing tools and voluntary participation. Additional data access may be required for scaling.

Project Lead(s): Marco Giacalone
Research Team: Independent research project led by Marco Giacalone, with the intention to progressively involve additional institutional collaborators from LICAIM and other professional domains as the study expands.
Start Date: August 2025
Expected Completion: April 2026
Last Updated: 10/31/2025