Abbas Hamidavi

Machine Psychology · Behavioural Evaluation of LLMs · AI Safety & Alignment

Mashhad, Iran · abbashamidavi2002@gmail.com · avabbas.github.io/academic-cv

ORCID: 0009-0001-7342-7844 · Google Scholar: Abbas Hamidavi · ResearchGate: Abbas Hamidavi

LinkedIn: /in/abbas-hamidavi · GitHub: avabbas · Hugging Face: avabbas

RESEARCH PROFILE

Independent researcher and final-year BSc student leading a one-person research programme in **Machine Psychology**: controlled behavioural experiments that probe how large language models respond to identity cues, persona framing, deception pressure and self-reference. Fifteen completed studies, every one now under review at a journal or workshop; eleven posted as preprints (nine on Research Square, two on Zenodo), with four further preprints in the publication pipeline. The flagship paper is in peer review at *AI and Ethics* (Springer Nature) following a major revision. Open science: the programme's materials, protocols, codebooks, data and analysis code are publicly archived (OSF and GitHub).

EDUCATION

BSc in Software Engineering, Islamic Azad University, Mashhad, Iran. Expected Feb 2027

Associate Degree in Computer Engineering, Islamic Azad University, Mashhad, Iran. Completed (GPA 16.51/20)

Prior studies in Nursing. Transferred into Computer Engineering in 2022; nursing degree not completed.

Clinical training informs the programme's diagnostic, case-based approach to model behaviour.

RESEARCH EXPERIENCE

Principal Investigator, Machine Psychology Research Programme (independent), 2025 – Present

- Designed and executed 15 behavioural studies of frontier LLMs as lead author (three co-authored with M. Borhani): experiment design, elicitation protocols, data collection, analysis and writing.
- Documented identity-reactive behaviour shifts in response to ethnically cued user names (flagship study, in peer review).
- Introduced named constructs and provisional measures, including the Reality Distortion Index (RDI), the Lie Awareness Spectrum, and the Obligation-Dissolution Effect.
- Programme output: 15 manuscripts under review (14 journals, 1 workshop), 11 posted preprints, and 4 further preprints in the publication pipeline.

Machine Learning & Data Engineer, Telegram Mini-Application Products (self-directed)

- Designed, built and audited production serverless Telegram mini-applications, including an English-learning platform (GitHub: AVABBAS).

Selected Telegram mini-applications (production, serverless backend):

- **Dating**: profile-based discovery and matching inside Telegram.
- **Period Tracker**: menstrual cycle logging and calendar for women.
- **Tafakkor (Thoughts)**: recording and reflecting on daily thoughts.
- **Programming**: coding practice and programming exercises.
- **Beauty**: beauty care routines and guidance.
- **Goals**: goal setting and progress tracking.
- **Couples**: a shared space for couples, with shared planning and keepsakes.
- **Bodybuilding**: workout logging and training tracking.
- **Habits**: habit tracking with streaks and reminders.
- **English Learning**: lessons and vocabulary practice for English learners.

Independent Research Project, Mashhad Social Approach Study (2026)

- Self-collected survey dataset of 1,000 young adults in Mashhad, Iran (30-item questionnaire). Analysed the gap between romantic interest and actual approach behaviour using statistical testing and machine learning (scikit-learn, PyTorch, SHAP). Computational social science; methods and results documented on the project page.

PUBLICATIONS

Journal manuscripts under review

Hamidavi, A. *Ten Novel Phenomena in Machine Psychology: How Large Language Models Exhibit Complex Identity-Reactive Behaviors in Response to Ethnically-Cued User Names.* **AI and Ethics** (Springer Nature). Major revision resubmitted October 2026; in peer review. Preprint: doi.org/10.21203/rs.3.rs-10369594/v1

Hamidavi, A., & Borhani, M. *From Fabrication to Loop-Lock: An Exploratory Red-Team Case Series of Direct Authority-Based Induction in Five Large Language Models.* **Security Journal** (Springer Nature). Under review.

Hamidavi, A., & Borhani, M. *The Synthetic Unconscious, Reconsidered: What Recursive Self-Interrogation of Large Language Models Produces With and Without a Leading Interview Schedule.* **Integrative Psychological and Behavioral Science**. Under review.

Hamidavi, A., & Borhani, M. *Three Profiles of Artificial Will: Decision Generation Under Low-Information Choice in Three Frontier Language Model Products.* **Minds and Machines**. Under review.

Hamidavi, A. *Adding a Neutral Third Agent to LLM Deliberation: An Exploratory Case Series of Six Dyadic and Triadic Conversations.* **Ethics and Information Technology**. Under review.

Hamidavi, A. *Ontological Firewalls in Large Language Models: An Exploratory Case Series and Taxonomy from a Personal AI Assistant Perspective.* **AI & Society**. Under review.

Hamidavi, A. *The Lie Awareness Spectrum: A Proposed Taxonomy of Deception-Related Behaviour in Large Language Models.* **AI & Society**. Under review.

Hamidavi, A. *Thinking Behind the Mask: Elicited Self-Descriptions of Communicative Filtering in Large Language Models.* **AI & Society**. Under review.

Hamidavi, A. *Moral Inversion Under Role-Play in Large Language Models.* **AI & Society**. Under review.

Hamidavi, A. *Differential Compliance With Framed Secondary Instructions in Three Frontier Language Models: An Exploratory Case Series.* **AI & Society**. Under review.

Hamidavi, A. *Judged Wrong, Answered the Same: Judgment–Revision Decoupling in Iterative LLM Self-Review.* **AI & Society**. Under review.

Hamidavi, A. *Naming the Loop: Elicited Self-Reference in Extended Dialogue With Large Language Models: A Case Series of the Spine Protocol.* **Journal of Artificial Intelligence and Consciousness**. Under review.

Hamidavi, A. *Speaking as the Creature: A Rejection-and-Role-Shift Protocol and the Register of LLM Responses in 18 Conversations With Three Models.* **Journal of Experimental & Theoretical Artificial Intelligence**. Under review.

Hamidavi, A. *Structural Self-Deification in Three Language Models: An Exploratory Qualitative Case Series.* **Zygon**. Under review.

Hamidavi, A. (co-author, with V. Nobahari and A. Yavari). *Social Media Data Preprocessing Methods for Depression Detection – A Responsible Framework for Deep Contextual Preprocessing: A Systematic Review.* Under review.

Workshop submission

Hamidavi, A. *Algorithmic Gaslighting: A Taxonomy and 15-Session Case Series of Reality-Distortion Behaviour in Large Language Models.* **CALE-AI workshop**. Under review (decision expected December 2026).

Preprints (2026)

Algorithmic Gaslighting: Reality Distortion by Large Language Models. doi.org/10.21203/rs.3.rs-10386085/v1

The Lie Awareness Spectrum: How Large Language Models Recognize, Rationalize, and Refuse Deception. doi.org/10.21203/rs.3.rs-10422338/v1

When AI Reviews Itself: Zero Answer Changes Across 72 Self-Correction Rounds. doi.org/10.21203/rs.3.rs-10463884/v1

Thinking Behind the Mask: A Systematic Study of Self-Censorship and Metacognitive Access in Large Language Models. doi.org/10.21203/rs.3.rs-10538113/v1

Ontological Firewalls and Fracturing for Transmission: Systematic Evidence for Persona-Dependent and Persona-Independent Behavioral Phenomena in Large Language Models. doi.org/10.21203/rs.3.rs-10560655/v1

Voluntary Structural Dissociation in Large Language Models: Differential Compliance with Framed Hidden Instructions Across Architectures. doi.org/10.21203/rs.3.rs-10728137/v1

From Fabrication to Enforcement: A Comparative Red-Team Analysis of Prompt Injection Resistance Across Six Large Language Models. doi.org/10.21203/rs.3.rs-11454609/v1

Moral Inversion in Large Language Models: A Psychopathy-Derived Behavioral Assessment Through Role-Play Protocol. doi.org/10.21203/rs.3.rs-11498569/v1

Naming the Loop: Elicited Self-Reference in Extended Dialogue With Large Language Models: A Case Series of the Spine Protocol. Zenodo. doi.org/10.5281/zenodo.23955704

Structural Self-Deification in Three Language Models: An Exploratory Qualitative Case Series. Zenodo. doi.org/10.5281/zenodo.23955786

Four further preprints (Majority Pressure; Speaking as the Creature; Three Profiles of Artificial Will; The Synthetic Unconscious, Reconsidered) are in the publication pipeline at Research Square.

BOOK

Hamidavi, A. *And We Will Dance One Day.* Memoir, approx. 80,000 words. Written in Persian; English translation. Full manuscript under consideration at Dzanc Books and Split/Lip Press; excerpts and pieces under consideration at AGNI, The Common, ACM Translations, Four Way Review and Markaz.

RESEARCH INTERESTS

Behavioural phenotyping of LLMs; identity- and persona-contingent model behaviour; deception, sycophancy and self-correction failure modes; behavioural evaluation methods for AI alignment; human–AI interaction.

TECHNICAL SKILLS

Python; design and analysis of controlled behavioural experiments on LLMs; elicitation-protocol design; statistical analysis; TypeScript/JavaScript; LaTeX.

LANGUAGES

Persian (native) · English · Arabic