Assessing Moral Development in AI-Based Introspective Assistants: Challenges and Strategies

Majid Gholami [*] | **Behrooz Minaei Bidgoli** [**] | **Hadi Hossenkhani**[***]

Abstract

The emergence of Large Language Models (LLMs) has opened new horizons for intelligent "introspective assistants" (Mohâsebe-ye Nafs), yet the abstract nature of moral concepts complicates their computational assessment. This research extracts, analyzes, and classifies the technical and conceptual challenges of evaluating morality within such assistants and formulates theoretical strategies to manage them. Findings are structured into a four-part framework: 1) Conceptual Translation Challenges, such as operationalizing intention and lacking ethical ontologies; 2) Data-Centric Challenges, including reliance on digital footprints, scarce standardized datasets, and cultural bias; 3) Algorithmic Logic Challenges, like model instability, the black box problem, and sensitivity to phrasing; 4) Interactive and Dynamism Challenges, encompassing flawed feedback loops and temporal misattribution error. The study concludes that overcoming these obstacles requires moving beyond purely statistical approaches. Proposed strategies involve developing multidimensional assessment indices, designing hybrid architectures that integrate explainable AI, causal inference, and formal ontologies based on Islamic ethics, and shifting from passive observation to active, user-centric self-evaluation.

Keywords

Introspection (Mohâsebe-ye Nafs), Moral Assessment, Large Language Models (LLMs), Artificial Intelligence, Intelligent Introspective Assistants.

[*] Ph.D. Candidate in Islamic Teaching, Imam Khomeini Educational and Research Institute, Qom, Iran. **(Corresponding Author)** | yabager248@gmail.com

[**] Professor, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran. | b_minaei@iust.ac.ir

[***] Assistant Professor, Department of Ethics, Imam Khomeini Educational and Research Institute, Qom, Iran. | hkh@iki.ac.ir

5

Abstracts