MODSIM World 2026

Abstract Submission — Keian Finlay

Adaptive Human Systems Lab | University of Central Florida

Track: Autonomy

Deadline: May 11, 2026 | Submit: easychair.org/my/conference?conf=modsim20260

Personality Traits and Trust Recovery in Autonomous Recommendation Systems: A Computational Simulation Study

Authors: Keian Finlay, Ancuta Margondai, Anamaria Acevedo Diaz, Mustapha Mouloua

University of Central Florida, Orlando, FL 32816, USA

Abstract

As autonomous systems become increasingly embedded in operational environments, understanding how individual personality differences shape trust dynamics following AI errors is critical for human-autonomy teaming design. While prior research has examined how personality traits influence initial trust formation in human-machine interaction (Zhu et al., 2025), the mechanisms through which distinct personality profiles recover trust across sequential interactions following an AI error remain underspecified, and existing empirical studies have been limited to short, static interaction sequences that cannot model trust dynamics over extended operational periods. This study addresses that gap by presenting a Python-based agent-based simulation grounded in the Big Five personality framework and Lee and See's (2004) foundational model of trust in automation.

The simulation deploys 1,000 synthetic agents across five Big Five personality profiles, including Emotional Stability, Conscientiousness, Openness, Agreeableness, and Sociability, each parameterized with theoretically grounded values for baseline trust,

learning rate, error sensitivity, and recovery rate derived from Big Five and trust in automation literature. Agents complete 20 sequential decision rounds interacting with an autonomous recommendation system operating at 85% baseline accuracy. A deliberate AI error is introduced at round 10, and trust scores, compliance rates, and post-error recovery trajectories are tracked across three communication conditions: no explanation, explanation with reasoning, and high-confidence messaging.

Simulation results reveal meaningful and theoretically consistent gradients in post-error trust recovery across personality profiles. Emotionally stable agents recovered 97.0% of pre-error baseline trust by round 20, followed by conscientiousness (96.5%), openness (95.1%), agreeableness (92.8%), and sociability (87.9%), consistent with Lee and See's (2004) predictions for affect-stable operators and interpersonal trust schema theory respectively. One-way ANOVA confirmed significant personality differences in post-error trust trajectories ($F=2913.74$, $p < .001$). The explanation with reasoning condition significantly outperformed no explanation in post-error recovery ($t=22.77$, $p < .001$, Cohen's $d=0.322$), while high-confidence messaging produced a significant negative effect ($t=-27.71$, $p < .001$, Cohen's $d=-0.392$), deepening trust collapse at the error point, a pattern consistent with betrayal-based trust collapse documented in interpersonal trust literature.

This work makes three contributions to the M&S community. First, it presents a computationally efficient simulation framework for modeling personality-differentiated trust dynamics in autonomous systems, providing a scalable alternative to large-scale human participant studies. Second, to the authors' knowledge, it identifies the trust-deepening effect of high-confidence messaging as a previously underspecified failure mode in AI communication design, with direct implications for human-autonomy interface development. Third, it demonstrates that communication strategies must be tailored to operator personality profiles to optimize trust recovery following AI errors, providing actionable design parameters for autonomous recommendation systems in operationally consequential environments. Findings have direct applicability to defense human-autonomy teaming, clinical decision support, and any operational domain in which AI error recovery determines mission effectiveness.

Key Simulation Results

The following results were produced by the verified simulation (1,000 agents x 20 rounds x 3 communication conditions, 60,000 total observations):

Post-Error Trust Recovery by Personality Profile (No Explanation Condition):

Emotionally Stable: 97.0% of pre-error baseline recovered by round 20

Conscientiousness: 96.5% of pre-error baseline recovered by round 20

Openness: 95.1% of pre-error baseline recovered by round 20

Agreeableness: 92.8% of pre-error baseline recovered by round 20

Sociability: 87.9% of pre-error baseline recovered by round 20

Statistical Results:

One-way ANOVA (post-error trust ~ personality): F=2913.74, p<.001

T-test Explanation vs No Explanation: t=22.77, p<.001, Cohen's d=0.322

T-test High-Confidence vs No Explanation: t=-27.71, p<.001, Cohen's d=-0.392

Note: The negative Cohen's d for high-confidence messaging (-0.392) confirms that this communication strategy actively deepens trust collapse at the error point rather than supporting recovery. This is the key novel finding and should be highlighted in the presentation.

References

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big Five personality domains. Journal of Research in Personality, 37(6), 504-528.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., & Mara, M. (2022). Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior, 139, 107539. https://doi.org/10.1016/j.chb.2022.107539

Zhu, Y., Hua, G., Liu, X., Wang, C., & Tang, M. (2025). Trust in machines: How personality trait shapes static and dynamic trust across different human-machine interaction modalities. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2025.1538369