The Effect of Robot's Flaws on Different Personality Traits

Conscientiousness and **agreeableness** are linked to smaller trust losses after robot errors, with measurable effects in some studies. **Extraversion** shows mixed effects, and evidence is insufficient or inconsistent for **openness** and **neuroticism**.

Conscientiousness and trust

Conscientiousness, associated with organization and diligence, tends to have a protective effect, correlating with less problematic reactions and potentially more patience or critical evaluation rather than immediate distrust [6]. Conscientiousness consistently appears associated with greater willingness to continue trusting robots after failures, especially in high-stakes scenarios; the magnitude of this effect varies across studies. Where measured, conscientious participants showed significantly higher ongoing trust following robot errors in emergency-type tasks, though some task contexts report only small overall personality effects.

- **Direction**: positive — higher conscientiousness correlates with less trust loss in several studies [1].
- **Strength**: study-dependent — significant in an emergency-context experiment (moderate effect) but personality effects were reported as relatively small in a collaborative manufacturing setting [1] [2].
- **Why**: conscientious people tend to be more methodical, dutiful, and tolerant of temporary imperfection when they believe reliability or procedure can be restored; they may focus on task rules and performance trends rather than isolated failures (logical synthesis).

Agreeableness and trust

Agreeableness is repeatedly linked to greater trust or smaller trust loss after robot mistakes, and lower agreeableness appears in participants who show persistent disbelief or low trust. Multiple studies identify agreeableness as a predictor of higher baseline trust and greater tolerance for robot errors.

- **Direction**: positive — higher agreeableness correlates with reduced trust loss following errors [1] [3].

- **Strength**: evidence shows significant associations in some experiments and also features in clustering analyses distinguishing low-trust participants, indicating a meaningful effect size in observed samples [1] [3].
- **Why**: agreeable individuals are more cooperative and dispositionally trusting of others (including agents), so they are likelier to attribute errors charitably or give the robot benefit of the doubt rather than immediately withdrawing trust (logical synthesis).

Extraversion and trust

Extraversion shows mixed but interpretable relations with trust dynamics: lower extraversion is associated with a subgroup that exhibits persistent disbelief and lower trust after robot interactions. Overall corpus evidence is limited and somewhat context-dependent.

- **Direction**: mixed — lower extraversion has been associated with being a “disbeliever” (more distrustful) in one clustering study, implying higher extraversion may relate to greater resilience to trust loss in some tasks [3].
- **Strength**: small-to-moderate and variable across contexts; not consistently strong across all experiments [3] [2].
- **Why**: extraverted people may be more socially engaged and open to interacting with agents, producing more positive initial rapport and a higher likelihood to forgive or re-engage after errors; less extraverted individuals may be more reserved and quicker to withhold trust (logical synthesis).

Openness and trust

People scoring high on openness tend to retain trust even when the robot errors. Openness predicted higher overall trust and faster reaction times, but it did not amplify trust loss after mistakes. The supplied studies do not provide clear, consistent empirical findings tying openness to the magnitude of trust loss after robot failures. Evidence in the corpus is insufficient to state a reliable direction or strength. [4]

- **Evidence statement**: insufficient evidence in the supplied literature to determine a robust correlation for openness.
- **Possible mechanism**: conceptually, higher openness could reduce trust loss because open individuals tolerate novelty and technological variability, while low openness could increase sensitivity to unexpected errors; this is an inferential explanation not directly supported by the provided studies (logical synthesis).

Neuroticism and trust

Neuroticism, characterized by emotional instability, is strongly linked to negative responses in related contexts like problematic smartphone use, indicating that Theoretically, higher neuroticism (greater anxiety/negative affect) may lose trust more quickly when a system malfunctions. [6] Neuroticism predicts lower trust ratings and a stronger negative reaction to system failures; higher neuroticism magnifies distrust after a flaw [5]. However, there is no clear, consistent empirical result in the supplied corpus linking neuroticism to trust loss from robot errors, so the direction and strength are indeterminate based on these sources.

Relative trait impact summary

The most consistently observed traits linked to smaller trust loss are **agreeableness** and **conscientiousness**, which appear to increase tolerance toward robot errors in multiple studies [1] [3]. **Extraversion** shows mixed associations with distrust clusters and therefore has a moderate, context-dependent effect [3]. **Openness** and **neuroticism** lack clear, reproducible findings in the supplied corpus, so their effects remain indeterminate based on available evidence [2].

References

[1] A. Xu and G. Dudek, “Towards Modeling Real-Time Trust in Asymmetric Human–Robot Collaborations,” pp. 113–129, Jan. 2016, doi: 10.1007/978-3-319-28872-7_7.

[2] S. Sarkar, D. Araiza-Illan, and K. Eder, “Effects of Faults, Experience, and Personality on Trust in a Robot Co-Worker”, doi: 10.48550/arxiv.1703.02335.

[3] P. Aliasghari, M. Ghafurian, C. L. Nehaniv, and K. Dautenhahn, “How Do We Perceive Our Trainee Robots? Exploring the Impact of Robot Errors and Appearance When Performing Domestic Physical Tasks on Teachers’ Trust and Evaluations,” *ACM transactions on human-robot interaction*, vol. 12, pp. 1–41, Feb. 2023, doi: 10.1145/3582516.

[4] Oksanen, A., Savela, N., Latikka, R., & Koivula, A. (2020). Trust Toward Robots and Artificial Intelligence: An Experimental Approach to Human–Technology Interactions Online. Frontiers in Psychology, 11, 568256. https://doi.org/10.3389/fpsyg.2020.568256

[5] Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence. Heliyon, 6(8), e04572. https://doi.org/10.1016/j.heliyon.2020.e04572

[6] Liu, C., Ren, L., Rotaru, K., Liu, X., Li, K., Yang, W., Li, Y., Wei, X., Yücel, M., & Albertella, L. (2023). Bridging the links between Big Five personality traits and problematic

smartphone use: A network analysis. Journal of Behavioral Addictions, 12, 128 - 136. https://doi.org/10.1556/2006.2022.00093.