The Drawback of Insight: Detailed Explanations Can Reduce Agreement with XAI

Sabid Bin Habib Pias[*]
Indiana University
Indiana, USA

Taslima Akter
UC Irvine
California, USA

Alicia Freel
Indiana University
Indiana, USA

Donald Williamson
Ohio State University
Ohio, USA

Timothy Trammel
UC Davis
California, USA

Apu Kapadia
Indiana University
Indiana, USA

Abstract

With the emergence of Artificial Intelligence (AI)-based decision-making, explanations help increase new technology adoption through enhanced trust and reliability. However, our experimental study challenges the notion that every user universally values explanations. We argue that the agreement with AI suggestions, whether accompanied by explanations or not, is influenced by individual differences in personality traits and the users' comfort with technology. We found that people with higher neuroticism and lower technological comfort showed more agreement with the recommendations without explanations. As more users become exposed to eXplainable AI (XAI) and AI-based systems, we argue that the XAI design should not provide explanations for users with high neuroticism and low technology comfort. Prioritizing user personalities in XAI systems will help users become better collaborators of AI systems.

Author Keywords

eXplainable AI; Technological Comfort; Personality Traits

CCS Concepts

•**Human-centered computing** $\rightarrow$ **Empirical studies in collaborative and social computing; Human computer interaction (HCI);**

Introduction

As an attempt to make eXplainable AI (XAI) more understandable and user-friendly, studies have explored the efficacy of single-style text explanation in varying AI-based solutions [12, 24]. Research showed that users have a higher perception of trust and transparency with text-based explanations with adequate information [15, 8]. Moreover, people's acceptance of new information from either a human or an AI agent depends on their personality traits [16, 14, 11]. However, the acceptance of explanations by XAI-based systems based on the relationship between individual personality traits and the type of explanation presented to the user is understudied. Therefore, we seek to understand how users with different personalities accept explanations containing varying levels of information in an AI classifier system. In our study, we investigated the influence of three types of single-style text explanations: 1) no explanation, 2) placebic, and 3) meaningful explanations [8] based on participants' personality traits (neuroticism and conscientiousness) and technology comfort.

We designed an online user study (N = 224) with a simulated AI classifier interface to compare how participants with varying technological comfort, neuroticism, and conscientiousness agreed with the classifier with varying explanation types. We show that participants with low technology comfort and participants with higher neuroticism prefer prompts without any explanation, showing the need to evaluate user technology comfort and personality traits before deploying explanations in any XAI-based classifier or recommender.

Are Explanations Acceptable by Everyone?

As XAI-based systems become popular in daily and productive tasks, more people are introduced to various AI technologies with explanations [20]. However, the ultimate goal of explainable AI is to aid users, making it important to understand if the preference for explanations varies based on the unique personality traits of users. For this reason, we focus on participants' tendency to agree with different levels of information (no explanation, placebic, meaningful) based on their neuroticism, conscientiousness, and tech comfort.

Neuroticism denotes the tendency or disposition to experience negative emotional states. People with high neuroticism are more likely to respond poorly to environmental stress and suspect negative outcomes from an ordinary situation [2, 25]. Conscientiousness is about being careful, diligent, and thorough in daily activities. A person with high conscientiousness tends to pay attention to details and ensure that their deliverables are correct [18]. Barnett et al. have shown that higher neuroticism negatively affects technology use and higher conscientiousness influences a person to use technologies [1]. Therefore, we contend that individuals exhibiting higher levels of neuroticism will opt for reduced information to mitigate distractions, and participants with higher conscientiousness will lean towards increased information provided by XAI-based systems to enhance task accuracy. Technological comfort refers to the ease and confidence of a person in using technological tools. This enables them to understand the technology and critically evaluate the information provided by the system. Park et al. [17] posited that users' confidence in technology-provided information tends to be lower among individuals with limited technological competence than users with higher technological skills. Therefore, we hypothesize that participants with lower technological comfort will prefer minimal information in XAI-based systems. Based on previous findings that analyze the effect of neuroticism, conscientiousness, and tech comfort on user agreeability, [18, 1], we propose the following hypotheses in the context of XAI.
H1: Participants with lower tech comfort agree more with

Figure 1: Sample blurred image of a trophy

Sample Explanations

Correct Explanations

No exp.: This is a trophy.
Placebic: This is a trophy because it looks like a trophy.
Meaningful: This is a trophy because it has a gold-colored cup and a white square base.

Incorrect Explanations

No explanation: This is a medal.
Placebic: This is a medal because it looks like a medal.
Meaningful: This is a medal because it is circular, gold and hanging from a ribbon.

---

classifications *without explanations* compared to classifications with *meaningful explanations*

H2: Participants with higher neuroticism agree more with classifications *without explanations* compared to classifications with *meaningful explanations*

H3: Participants with higher conscientiousness agree more with classifications with *meaningful explanations* compared to classifications *without explanations*

Method

The study consisted of a norming study and an agreeability measurement study, both using a between-subject study design. In the norming study, 100 well-known object images were selected from the BOSS stimulus set [4], and Gaussian blur filters with various levels were applied to determine the optimal blur level (sigma 115) that balanced identifiability. Participants (N = 60) viewed 50 blurred images randomly from a pool of 61 images. They identified the object and briefly described the characteristics that helped their identification. We used the most frequently stated characteristics to generate meaningful explanations and the second most frequently labeled objects as the incorrect labels for the agreement measure study. This ensured that the meaningful explanations for both correct and incorrect labels were based on object characteristics that humans found useful in identifying distorted images. In the agreeability measurement study, participants (N=225) were randomly assigned to one of three explanation conditions (no explanation, placebic, and meaningful) as part of the design between subjects. They were asked to rate their agreement with the AI's classification of the image on a 7-point Likert scale. Among the 60 images shown to each participant, 48 (80%) were correctly labeled for the three explanation conditions. To ensure accurate balance, correct answers were counterbalanced using a modified Latin square design [19]. The study explored the relationship between personality, technology comfort, and participant agreement with the AI classifier for the three types of explanations. We used the Big 5 personality inventory (10-item version) [3] to assess the neuroticism and conscientiousness of the participants via the mean score of the corresponding items. To measure technology comfort we simply asked participants how much they agreed/disagreed with the following statement "Operating and using everyday technology is easy for me". Additionally, we incorporated the technology acceptance questionnaire (TAM) [7]. The participants were recruited through Prolific [1] and were all adults from the United States. We paid $7.00 to each participant for a completed study regardless of their response quality. The payment amount was calculated to be above $15/hour which surpasses the recommended minimum wage [22] in the study's location. Statistical analysis used a linear mixed-effects model, treating explanation type, technology comfort, neuroticism, and conscientiousness as fixed effects, and participant and images as random effects.

Findings

Effect of Technology Comfort and Explanation Type
Participants indicating lower level of comfort with new technology were more likely to agree with the AI classifier when there were no explanations, compared to the other two conditions in which the AI classifier presented meaningful and placebic explanations (Fig. 3). *This finding suggests that users low in technological comfort are less comfortable navigating 'new' technology that explains itself (supporting H1).*

Effect of Neuroticism and Explanation Type
From Table 1, it is clear that individuals with greater neuroticism were more likely to agree with the recommendations that lacked explanations, compared to the recommendations with either meaningful or placebo information. In con-
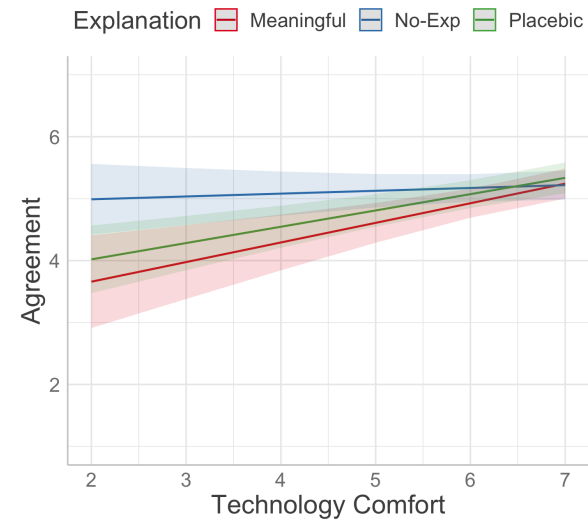
---

[1] https://www.prolific.com/
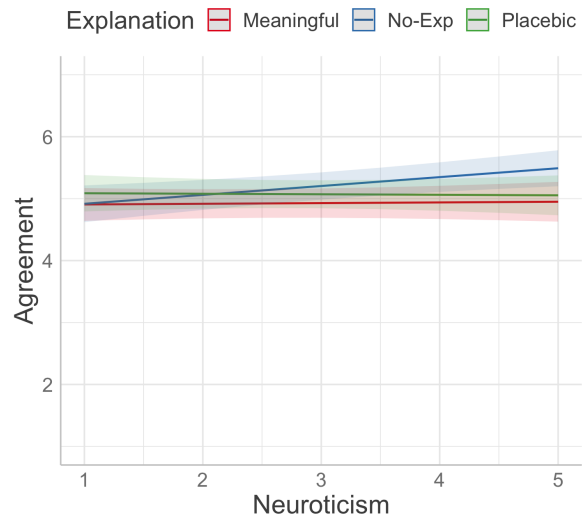
Figure 2: Relation of neuroticism and agreement to AI classifier

Figure 3: Relation of tech comfort and agreement to AI classifier

trast, participants with lower levels of neuroticism showed no differences in their agreement with recommendations based on the type of explanation provided. (Fig. 2). *The result indicates that participants with higher emotional volatility think that the classifier is more likely to be correct when the classifier provides less information compared to when the classifier provides more text in the form of meaningful and placebic explanations (supporting H2).*

Effect of Conscientiousness and Explanation Type
In contrast to our hypotheses, we did not find significant differences in participant agreement with the AI classifier presenting varying types of explanations (Table 1). This result suggests that conscientiousness may not directly affect the

reaction of the participant to the varying explanation types, failing to support H3.

Discussion
Increased agreement to classifications without explanation can provide insight for inclusive eXplainable AI (XAI) design, facilitating action-oriented nudges for improved user collaboration with fast-moving AI. Our findings show that individuals with higher neuroticism and low-tech comfort tend to prefer minimal or no explanations. The inclination of participants with lower comfort in technology towards preferring minimal information aligns with previous research indicating that decreased technological competence correlates with reduced reliance on technology, resulting in

Effects	Estimate	Std. Error
TC: Meaningful	0.27^{**}	0.1
TC: Placebic	0.21^{*}	0.08
Neuro: Meaningful	-0.13^{*}	0.06
Neuro: Placebic	-0.15^{*}	0.07
Conscience: Meaningful	-0.15	0.09
Conscience: Placebic	-0.06	0.1

Table 1: Linear Mixed Model results for Agreement with the AI classifier. (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$). TC = Tech Comfort, Neuro= Neuroticism, Conscience= Conscientiousness

diminished trust in information delivered by new technological platforms [17]. Furthermore, excessive information can overwhelm certain personalities and increase perceived complexity [13], reducing efficacy for people with higher neuroticism. Moreover, perceived technological insecurity, particularly among people with high neuroticism and low-tech comfort, may contribute to a preference for minimal explanations from AI classifiers [6, 23, 10]. Psychologically informed approaches to AI-based systems are recommended, highlighting the importance of personalizing explanations based on user traits to improve acceptance and adaptation to AI technologies [9]. For users with low technological comfort, providing personalized support and training resources tailored to their existing comfort levels is crucial to enhance their technology and AI comfort.

Our lack of supporting evidence for effects between conscientiousness and the preference for meaningful explanations fails to support Chiou et al.'s finding that people with high internal locus of control (LoC) are comfortable with robots exhibiting more information and control for navigation tasks [5]. LoC refers to the belief of people that their actions and behaviors affect the outcomes of events in their lives [21]. Interestingly, participants with higher conscientiousness do not have a higher preference for meaningful explanations, that provide more control to the participants.

Designing XAI systems considering user personality and preferences is crucial for ensuring user satisfaction and XAI systems' effectiveness. Different users have varying thresholds for information complexity and detail, influenced by their personality traits, cognitive styles, and context-specific needs. User acceptance and trust in XAI systems can be enhanced by incorporating user-profiles and allowing for adjustable explanation levels. This approach respects individual user preferences and improves the overall user experience by providing information in explanations as the user requires. Such flexibility ensures that XAI systems are not just tools for transparency, but are also adaptable interfaces that enable users to engage with AI on their own terms.

Conclusion
The interplay between user personality traits, technological comfort, and explanation level presents both challenges and opportunities to create more effective and engaging XAI designs. Certain user traits like higher neuroticism and lower technological comfort, need a "less is more" approach to ensure accessible and inclusive XAI design. By paying attention to the diverse needs and preferences of users, XAI designers can develop systems that are more welcoming, easier to use, and conducive to human-AI collaboration.

Acknowledgements

REFERENCES

[1] Tim Barnett, Allison W Pearson, Rodney Pearson, and Franz W Kellermanns. 2015. Five-factor model personality traits as predictors of perceived and actual usage of technology. *European Journal of Information Systems* 24, 4 (July 2015), 374–390. DOI: `http://dx.doi.org/10.1057/ejis.2014.10`

[2] Murray R. Barrick and Michael K. Mount. 1991. The Big Five Personality Dimensions and Job Performance: A Meta-Analysis. *Personnel Psychology* 44, 1 (1991), 1–26. DOI:`http://dx.doi.org/10.1111/j.1744-6570.1991.tb00688.x` _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1744-6570.1991.tb00688.x.

[3] Oliver P. John Beatrice Rammstedt. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212.

[4] Mathieu B Brodeur, Emmanuelle Dionne-Dostie, Tina Montreuil, and Martin Lepage. 2010. The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PloS one* 5, 5 (2010), e10773.

[5] Manolis Chiou, Faye McCabe, Markella Grigoriou, and Rustam Stolkin. 2021. Trust, Shared Understanding and Locus of Control in Mixed-Initiative Robotic Systems. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. 684–691. DOI:`http://dx.doi.org/10.1109/RO-MAN50785.2021.9515476` ISSN: 1944-9437.

[6] Kevin Craig, Jason Bennett Thatcher, and Varun Grover. 2019. The IT Identity Threat: A Conceptual Definition and Operational Measure. *Journal of Management Information Systems* 36, 1 (Jan. 2019), 259–288. DOI: `http://dx.doi.org/10.1080/07421222.2018.1550561` Publisher: Routledge _eprint: https://doi.org/10.1080/07421222.2018.1550561.

[7] Fred D Davis and others. 1989. Technology acceptance model: TAM. *Al-Suqri, MN, Al-Aufi, AS: Information Seeking Behavior and Technology Adoption* 205 (1989), 219.

[8] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The Impact of Placebic Explanations on Trust in Intelligent Systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. DOI: `http://dx.doi.org/10.1145/3290607.3312787`

[9] Davide Gentile, Greg Jamieson, and Birsen Donmez. 2021. Evaluating human understanding in XAI systems. In *ACM CHI XCXAI Workshop*.

[10] Andrea L Guzman and Seth C Lewis. 2020. Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society* 22, 1 (Jan. 2020), 70–86. DOI: `http://dx.doi.org/10.1177/1461444819858691` Publisher: SAGE Publications.

[11] Feridun Kaya, Fatih Aydin, Astrid Schepman, Paul Rodway, Okan Yetişensoy, and Meva Demir Kaya. 2024. The Roles of Personality Traits, AI Anxiety, and Demographic Factors in Attitudes toward Artificial Intelligence. *International Journal of Human–Computer Interaction* 40, 2 (Jan. 2024), 497–514. DOI: `http://dx.doi.org/10.1080/10447318.2022.2151730` Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2151730.

[12] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 379–390.

[13] Jona Linde, Daniel Gietl, Joep Sonnemans, and Jan Tuinstra. 2023. The effect of quantity and quality of information in strategy tournaments. *Journal of Economic Behavior & Organization* 211 (July 2023), 305–323. DOI: `http://dx.doi.org/10.1016/j.jebo.2023.04.024`

[14] Lorilee A. Medders and Sarah K. Baisley. 2017. Personality traits and thinking dispositions: Individual differences call for 'smart' messaging. *Journal of Business Continuity & Emergency Planning* 10, 4 (Jan. 2017), 367–383.

[15] Jeroen Ooge, Shotallo Kato, and Katrien Verbert. 2022. Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces*. 93–105.

[16] Jiyoung Park and Sang Woo. 2022. Who Likes Artificial Intelligence? Personality Predictors of Attitudes toward Artificial Intelligence. *The Journal of Psychology* 156 (Jan. 2022), 1–27. DOI: `http://dx.doi.org/10.1080/00223980.2021.2012109`

[17] Soyoung Park, Joseph D. Straubhaar, and Sharon L. Strover. 2019. Exploring Ambivalence in Technological Embeddedness: The Role of Technological Competence and Dependence in the Information Gap. *Journal of Broadcasting & Electronic Media* 63, 3 (July 2019), 433–454. DOI: `http://dx.doi.org/10.1080/08838151.2019.1653105` Publisher: Routledge _eprint: https://doi.org/10.1080/08838151.2019.1653105.

[18] Brent W. Roberts, Carl Lejuez, Robert F. Krueger, Jessica M. Richards, and Patrick L. Hill. 2014. What is conscientiousness and how can it be assessed? *Developmental Psychology* 50, 5 (2014), 1315–1330. DOI:`http://dx.doi.org/10.1037/a0031109` Place: US Publisher: American Psychological Association.

[19] B Rojas and RF White. 1957. The modified Latin square. *Journal of the Royal Statistical Society: Series B (Methodological)* 19, 2 (1957), 305–317.

[20] Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023), 1–20. DOI: `http://dx.doi.org/10.1109/TPAMI.2023.3331846` Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[21] Julian B. Rotter. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied* 80, 1 (1966), 1–28. `DOI:` `http://dx.doi.org/10.1037/h0092976` Place: US Publisher: American Psychological Association.

[22] M Six Silberman, Bill Tomlinson, Rochelle LaPlante, Joel Ross, Lilly Irani, and Andrew Zaldivar. 2018. Responsible research with crowds: pay crowdworkers at least minimum wage. *Commun. ACM* 61, 3 (2018), 39–41.

[23] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 74–88. `DOI:` `http://dx.doi.org/10.1093/jcmc/zmz026`

[24] Ellen M. Voorhees. 2021. System Explanations: A Cautionary Tale. *NIST* (May 2021). `https://www.nist.gov/publications/` `system-explanations-cautionary-tale` Last Modified: 2023-02-14T11:02-05:00 Publisher: Ellen M. Voorhees.

[25] Thomas A. Widiger. 2009. Neuroticism. In *Handbook of individual differences in social behavior*. The Guilford Press, New York, NY, US, 129–146.