Proceedings of the 5th International Conference on Social Psychology and Humanity Studies
DOI: 10.54254/2753-7048/2027.38382

Research on the Influence of Agreeableness and Neuroticism of the Five Major Personality Traits on Susceptibility to Misinformation

Zhitian Feng

Shanghai Starriver Bilingual School, Shanghai, China
Wesley.feng@hotmail.com

Abstract. Social deduction games require people to identify the truth-tellers and liars based on limited and vague information. This exploratory study investigated the relationship between agreeableness, neuroticism and accuracy of deception judgment in a structured social-deduction task. Sixty-three participants aged 15-35 completed a 120-item Big Five personality inventory and then observed the materials from a Werewolf game before making five judgments, three hidden-role judgments and two truth-value judgments. The results showed that neuroticism was negatively correlated with the accuracy of judgment ($r = -0.294$, $p = .019$, 95% CI [-0.505, -0.050]), and agreeableness was not significantly related to accuracy ($r = -0.128$, $p = .317$, 95% CI [-0.364, 0.124]). Exploratory analysis of trait combinations also showed that the mean accuracy of the agreeableness and neuroticism groups was different. Therefore, it is proposed that neuroticism may be linked to differences in the judgement of deception among individuals, but agreeableness alone was not significantly associated with performance in the current task. As an exploratory correlational study, the results provide an initial indication of personality differences in deception judgment under the circumstances of structured social interaction.

Keywords: Big Five, neuroticism, agreeableness, deception detection, social deduction

1. Introduction

Deception is a type of social behaviour; at present, it is difficult to recognise false information. Due to the lack of information and indistinct social signals, people make presumptions. Previous studies on lie detection have shown that, without the help of devices, people can only detect some deception at a slightly higher rate than chance, and they may be subject to other factors [1]. The above defects show that the reasons for variations in the precision of deception judgment need to be examined individually.

Social deception games provide a convenient and relatively controlled environment for this research. In the game "Werewolf", players hide their true identities and then talk about everything in the game. Compared with real-life situations of misinformation, the information environment and task rules in "Werewolf" are more organised and easier to observe. Thus, the game offers a task-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7048/2027.38382

oriented scenario for research on individual differences in deception judgment and does not directly link game performance with vulnerability to misinformation in daily life [2].

Personality is a possible origin of individual differences. The "Big Five" personality model is often used to study the connection between traits of personality and cognitive or social behaviour [3, 4]. This study anticipates that "Agreeableness" and "Neuroticism" are associated with decision-making processes in situations characterized by social uncertainty. Agreeableness is linked to trust, cooperation, and prosocial behavior, whereas Neuroticism is associated with negative emotions and emotional instability [3, 4]. Previous research suggests that high Agreeableness is associated with a greater willingness to meet the needs of others, while high Neuroticism is linked to a greater susceptibility to misinformation [5, 6]. However, few studies have investigated whether there is a link between these specific personality traits and the accuracy of deception judgments within structured social deception contexts. Therefore, this study examines the relationships between Agreeableness, Neuroticism, and the accuracy of deception judgment among participants after they complete a "Werewolf" game task, and determines whether different combinations of these two traits are associated with variations in judgment accuracy. Based on these objectives, the following hypotheses are proposed:

H1: Higher levels of Agreeableness are associated with lower accuracy in deception detection (a negative correlation).

H2: Higher levels of Neuroticism are associated with lower accuracy in deception judgment.

H3: Different combinations of Agreeableness and Neuroticism are associated with differences in deception judgment accuracy.

2. Method

2.1. Research design and procedure

A cross-sectional individual-differences design and a standardised online-judgment task were employed in this study. Sixty-three participants aged 15-35 years were recruited from the public. Participants were informed of the study's content and the right to withdraw. In all, there were three sessions. In the preparation stage, participants completed a personality trait questionnaire and edited a 30-minute video of the Werewolf game. Participants watched a video in the experiment and judged the identities of the werewolves and whether the statements presented were true or false. The reason for the judgment was also recorded. Organise and analyse the collected data in the final stage using Python.

2.2. Experimental materials and measurement tools

Personality trait measurement. The Chinese version of the internationally used Big Five Inventory from Big5PersonalityTest.com was employed, focusing on agreeableness and neuroticism. The 120-item scale, an official translation of the widely used English version, uses a 5-point Likert rating and has good reliability and validity [7]. The scale scores served as the indicators of agreeableness and neuroticism.

Experimental stimuli. The core material was a standardized 30-minute Werewolf game video, with scripts designed to manipulate exposure to true, false, and mixed misleading speech. All instructions and procedures were standardized and pre-tested for clarity and neutrality.

Behavioral indicators. Core performance metrics included five judgment-accuracy items, the number of times participants were misled, and their reasoning in the Werewolf task. Only when a

DOI: 10.54254/2753-7048/2027.38382

judgment and its corresponding reasoning process aligned was it counted as 1 mark.

2.3. Data collection and analysis methods

Data were collected on standardized personality scores (agreeableness and neuroticism) and judgment scores (marked 1–5). The analysis plan included descriptive statistics, Pearson correlation analysis, linear regression trend analysis, quadrant-group comparison, and hypothesis testing. All data were processed using Python.

3. Results

3.1. Descriptive statistics and preliminary analysis

A total of 63 participants (male and female, aged 15–35) completed the experiment. Scores on the Big Five traits approximated a normal distribution. Agreeableness (M = 67.81, SD = 24.23) and neuroticism (M = 63.06, SD = 22.35), the core variables, were moderately high and moderate, respectively. Participants' average judgment accuracy in the Werewolf game was 69.2% (SD = 18.9%), with 16 of 63 participants (25.4%) achieving 100% accuracy.

Table 1. Descriptive statistics of key variables (N = 63)

Dimension	Mean (M)	Standard Deviation (SD)
Agreeableness	67.81	24.23
Neuroticism	63.06	22.35
Judgment accuracy (%)	69.2	18.9

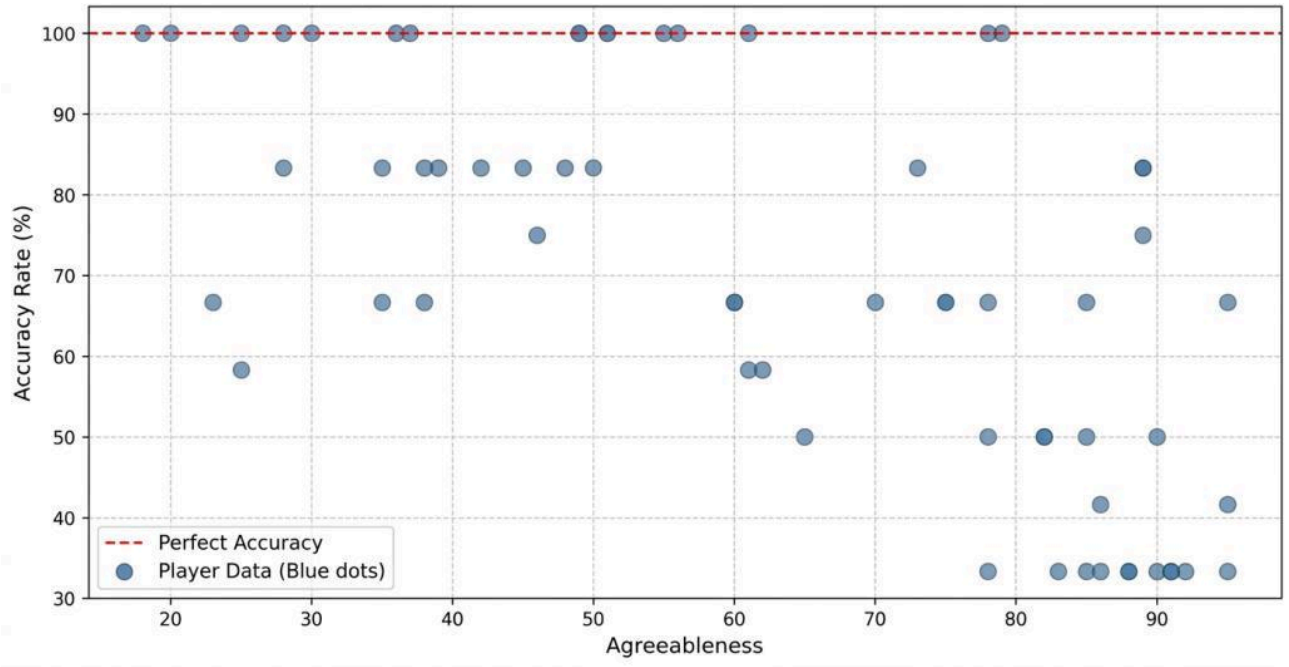


Figure 1. Agreeableness and wolf detection accuracy

DOI: 10.54254/2753-7048/2027.38382

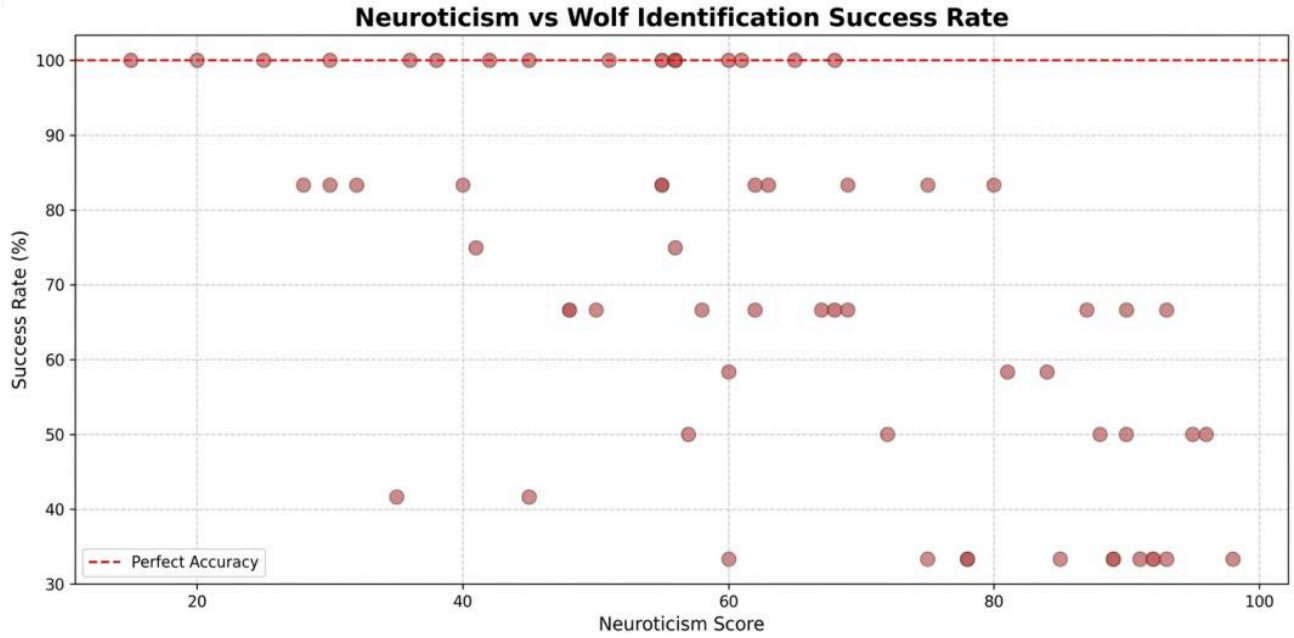


Figure 2. Neuroticism and wolf-detection success rate

As shown in Figures 1 and 2, judgment accuracy generally declined as neuroticism increased, whereas the relationship between agreeableness and accuracy was less systematic. For agreeableness, 9 of the 16 participants with perfect scores had low agreeableness (<50); for neuroticism, 6 of the 16 participants with perfect scores had low neuroticism (<40).

3.2. Correlation analysis results

Pearson correlation analysis preliminarily examined the relationship between personality traits and susceptibility to misinformation. Agreeableness was weakly and non-significantly negatively correlated with judgment accuracy, $r = -0.128$, $p = .317$: more agreeable (kind, cooperative, trusting) individuals showed slightly lower accuracy, but this trend may be due to chance. Neuroticism showed a moderate and significant negative correlation with accuracy, $r = -0.294$, $p = .019$: emotionally unstable, anxious individuals had lower accuracy, and this relationship was unlikely to be due to chance. In short, neuroticism significantly negatively predicted accuracy—emotionally stable, calm players performed better—whereas agreeableness had no significant effect. "Kind individuals" were not necessarily more easily deceived, at least in this sample.

Comprehensive analysis of two-dimensional personality traits. To further examine how the two traits jointly relate to judgment accuracy, we plotted a "personality coordinate plane" with agreeableness on the X-axis and neuroticism on the Y-axis, mapped each player's accuracy as point colour and size (heatmap/bubble-chart logic), and analysed average accuracy across personality quadrants. Using the means of agreeableness (M = 67.81) and neuroticism (M = 63.06) as cut-offs, the 63 players were divided into four quadrants (Table 2).

Table 2. Quadrant definitions based on mean cut-offs

Quadrant	Agreeableness	Neuroticism	Trait Description
Q1	High (≥67.81)	High (≥63.06)	Kind but emotionally unstable — "Anxious Nice Person"
Q2	Low (<67.81)	High (≥63.06)	Skeptical and emotional — "Irritable Skeptic"

DOI: 10.54254/2753-7048/2027.38382

Table 2. (continued)

Q3	Low (<67.81)	Low (<63.06)	Calm and uncooperative — "Rational Detached"
Q4	High (≥67.81)	Low (<63.06)	Kind and stable — "Buddhist Master"

After classifying players into quadrants and calculating the average accuracy (ACC) per quadrant:

Table 3. Average accuracy by personality quadrant

Quadrant	Number of People	Average Accuracy (ACC)	Trait Summary
Q1 (High A, High N)	17	0.625	Kind but anxious; conservative judgments; often release werewolves
Q2 (Low A, High N)	14	0.609	Suspicious and emotional; prone to wrongfully accusing villagers and confusing logic
Q3 (Low A, Low N)	16	0.713	Calm and rational; most accurate; few missed werewolves or false accusations
Q4 (High A, Low N)	16	0.698	Kind and stable; less likely to suspect, but clear logic

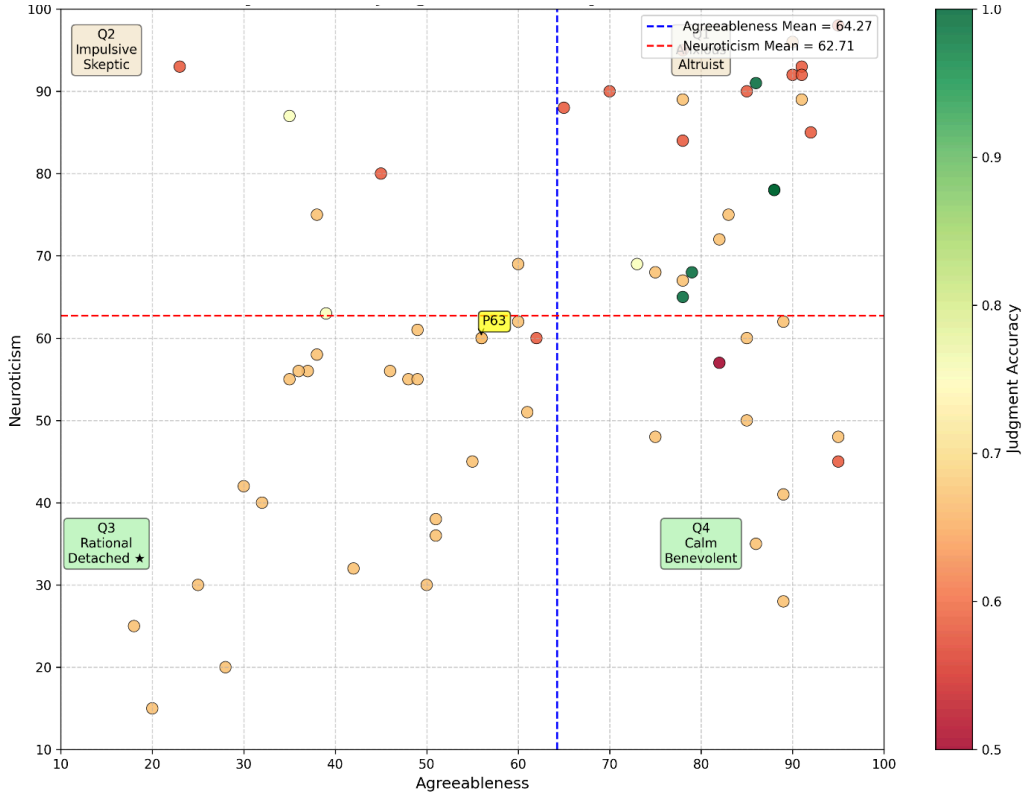


Figure 3. Personality coordinate plane: agreeableness × neuroticism and judgment accuracy

Q3 and Q4 (the low-neuroticism groups) had higher average accuracy than the high-neuroticism groups (Q1 and Q2), revalidating the Pearson correlation result—emotional stability is critical. The most accurate group was Q3 ("Rational Detached"), with an average accuracy of 71.3%, whereas the least accurate group was Q2 ("Irritable Skeptic"), with 60.9%. Each 1-unit increase in neuroticism

DOI: 10.54254/2753-7048/2027.38382

was associated with a decrease in accuracy of approximately 0.003 ($p < .05$), whereas each 1-unit increase in agreeableness was associated with a decrease of approximately 0.001 ($p > .05$). Thus, on the personality plane, accuracy increased from the top-left to the bottom-right, with the optimal region in the low-neuroticism half-plane. In summary, correlation and regression analyses consistently show that neuroticism is an important personality factor predicting susceptibility to misinformation, providing strong empirical support for Hypothesis 2.

4. Discussion

4.1. Key research findings

This study examined the associations between agreeableness, neuroticism, and deception judgment accuracy in a social-deduction task. Consistent with previous research showing that human deception judgments are generally limited in accuracy and often rely on imperfect cues [8, 9], the present findings suggest that personality may be relevant to individual differences in judgment performance.

H1 (agreeableness) was not supported. Agreeableness was not significantly associated with judgment accuracy ($r = −0.128$, $p = 0.317$). Although participants with higher agreeableness showed a slight tendency toward lower accuracy, the association was weak and not statistically significant. This finding suggests that agreeableness alone may not be a strong determinant of deception judgment accuracy in the present task.

H2 (neuroticism) was supported. Neuroticism was significantly negatively associated with judgment accuracy ($r = −0.294$, $p = 0.019$). Participants with higher neuroticism tended to achieve lower accuracy in the deception judgment task. This finding suggests that individual differences in emotional reactivity may be relevant to judgment performance under uncertain and socially deceptive conditions.

H3 was partially supported at the exploratory level. The quadrant analysis showed differences in mean accuracy across combinations of agreeableness and neuroticism. The low-agreeableness–low-neuroticism group achieved the highest mean accuracy (71.3%), whereas the low-agreeableness–high-neuroticism group achieved the lowest (60.9%). These descriptive differences suggest that neuroticism may be particularly relevant when considering personality combinations. However, the quadrant analysis does not directly establish an interaction effect between the two traits, and these patterns should therefore be interpreted as exploratory.Error types. In terms of judgment errors, misjudging villagers as werewolves (30.1%) and werewolves as villagers (38.5%) differed significantly. Players were more likely to miss werewolves (38.5%) than to falsely accuse villagers (30.1%), showing a conservative bias—preferring to spare suspects rather than wrongly eliminate teammates. However, nearly one-third of villagers were falsely accused, indicating that "aggressive werewolf" strategies effectively misled some players. Such a tendency to let offenders go rather than accuse good people may, in real-world deception contexts, make wrongdoing easier to carry out.

Overall, agreeableness may be bidirectional in its effect, and its mechanism deserves further study. Neuroticism-driven emotionality may lead to more irrational judgments, and when combined with low agreeableness, judgments may become more extreme. Cognitive dispositions such as these may be trainable through anti-deception practice.

DOI: 10.54254/2753-7048/2027.38382

4.2. Interpretation and theoretical significance

Neuroticism is negatively related to the accuracy of deception judgement, and therefore, emotions may be involved in the decision-making process under uncertainty. In the Werewolf task, the participants had to judge the ambiguous information and react to social pressure and conflicting statements. Earlier studies have found that one's emotions can affect the choice and risk perception of individuals [10, 11]. Therefore, a person with higher neuroticism is more likely to have unstable judgments in socially uncertain situations. However, this interpretation is not concrete because the current study has not measured the participants' emotions or thoughts during the experiment directly.

Agreeableness is not affected by an increase in the level of judgment. Agreeableness is linked to trust and cooperation with others and prosocial behaviour; however, these traits were not significantly associated with a lower error rate in this experiment. Therefore, a tendency to trust others may not lead to a decreased perception of falsity when many sources of information are provided and an explicit role is designated for the evaluation. Therefore, it can be inferred that the correlation between personality and judgement of deception varies according to the specific task and social environment. Exploratory analysis of personality combinations suggests that, to fully understand the differences in people's judgment of social deception, several dimensions of personality should be taken into account.

4.3. Practical implications and application value

The conclusions provide a solid scientific basis for precision anti-fraud and information-literacy education. In education and publicity, anti-fraud strategies can shift from "one-size-fits-all" to "precision-targeted outreach": for high-agreeableness groups, educational content should focus on critical-thinking training and "healthy skepticism," using role-playing to teach maintaining kindness while establishing psychological boundaries; for high-neuroticism groups, the focus should be on emotional-management skills and structured information-verification methods to help them activate rational checks during emotional fluctuations. Personality-trait screening can serve as an effective tool to identify high-risk individuals for proactive, differentiated anti-fraud training in campuses and enterprises. In public safety and financial-risk prevention, law-enforcement organs can work with mental health institutions to build community early warning systems for high-risk groups. Financial institutions can add a personality-assessment model to the customer risk-identification system to perform enhanced verification or multi-level prompts for users with certain characteristics during transactions. The research suggests a multi-level anti-fraud system for social governance. Increase the capacity of the people and improve the overall information environment. Platforms can use the above information to optimise the content-review algorithm, reduce the spread of emotional and false information, and protect vulnerable groups at the source.

5. Conclusion

This research experimentally explored the impact of agreeableness and neuroticism (from the Big Five personality traits) on susceptibility to misinformation. The results partially validate the research hypotheses, providing new empirical evidence for understanding the relationship between personality traits and information processing. In misleading speech, neuroticism explains erroneous decisions better than agreeableness. Kindness itself is not an origin of misjudgment; emotional stability is more crucial. In summary, this research provides multi-level solutions for addressing misleading information, helping individuals improve information discernment and enabling

DOI: 10.54254/2753-7048/2027.38382

institutions to develop evidence-based prevention strategies. As the information environment grows more complex, personality-based differentiated strategies will become critical to enhancing overall information literacy.

Despite meaningful findings, this study has several limitations. First, the sample size was small (n = 63), and covariates such as age, gender, and game experience were not considered. Second, we cannot prove that all participants paid attention to the video. In addition, the differences between the Werewolf game and real-life fraud prevent the findings from being directly applied to the anti-fraud field. Finally, theoretically, this study focused only on agreeableness and neuroticism. Future research should develop integrated models including all Big Five dimensions, cognitive abilities, and media literacy, and should examine more complex personality configurations to reveal more comprehensively whether people high in agreeableness are affected under social pressure.

References

[1] Denault, V., Talwar, V., Plusquellec, P., & Larivière, V. (2022). On deception and lying: An overview of over 100 years of social science research. *Applied Cognitive Psychology, 36*(4), 805–819. https://doi.org/10.1002/acp.3971

[2] Granic, I., Lobel, A., & Engels, R. C. (2014). The benefits of playing video games. *American Psychologist, 69*(1), 66–78.

[3] Vassend, O., & Skrondal, A. (2011). The NEO personality inventory revised (NEO-PI-R): Exploring the measurement structure and variants of the five-factor model. *Personality and Individual Differences, 50*(8), 1300–1304.

[4] Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48*(1), 26–34.

[5] Whitty, M. T. (2019). Cyberscams: The role of personality and the internet in facilitating victimization. *Cyberpsychology, Behavior, and Social Networking, 22*(5), 317–323.

[6] Anderson, K. B., & Mutz, D. C. (2020). Fear and loathing across the political divide: How emotion drives misinformation sharing. *Proceedings of the National Academy of Sciences, 117*(47), 29589–29598.

[7] John, O. P., Naumann, L. P., & Soto, C. J. (2008). Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). The Guilford Press.

[8] Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review, 10*(3), 214–234.

[9] Hartwig, M., & Bond, C. F. (2011). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin, 137*(4), 643–659.

[10] Rolls, E. T. (2023). Emotion, motivation, decision-making, the orbitofrontal cortex, anterior cingulate cortex, and the amygdala. *Brain Structure and Function, 228*(5), 1201–1257. George,

[11] J. M., & Dane, E. (2016). Affect, emotion, and decision making. *Organizational Behavior and Human Decision Processes, 136*, 47–55.