

# Predicting Trust Dynamics With Personal Characteristics

Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2024, Vol. 68(1) 310–316  
Copyright © 2024 Human Factors and Ergonomics Society  
DOI: 10.1177/10711813241260383  
journals.sagepub.com/home/pro



Hyesun Chung<sup>1</sup> and X. Jessie Yang<sup>1</sup>

## Abstract

Previous research into trust dynamics in human-autonomy interaction has demonstrated that individuals exhibit specific patterns of trust when interacting repeatedly with automated systems. Moreover, people with different types of trust dynamics have been shown to differ across seven personal characteristic dimensions: masculinity, positive affect, extraversion, neuroticism, intellect, performance expectancy, and high expectations. In this study, we develop classification models aimed at predicting an individual's trust dynamics type—categorized as Bayesian decision-maker, disbeliever, or oscillator—based on these key dimensions. We employed multiple classification algorithms including the random forest classifier, multinomial logistic regression, Support Vector Machine, XGBoost, and Naive Bayes, and conducted a comparative evaluation of their performance. The results indicate that personal characteristics can effectively predict the type of trust dynamics, achieving an accuracy rate of 73.1%, and a weighted average *F1* score of 0.64. This study underscores the predictive power of personal traits in the context of human-autonomy interaction.

## Keywords

trust dynamics, personal characteristics, clustering, classification model, human-autonomy interaction

## Introduction

Trust in automation/autonomy<sup>1</sup>, defined as “the attitude that an automated or autonomous agent will help achieve an individual's goals in situations characterized by uncertainty and vulnerability” (Lee & See, 2004), has attracted increasing attention across various fields, including human-computer interaction, human factors, and engineering. Numerous factors have been recognized as antecedents of trust, encompassing system-related aspects like automation reliability, level and stage of automation, and transparency, personal factors such as personality, culture, and prior experiences, and environmental factors including task type and complexity (Hancock et al., 2011, 2021; Hoff & Bashir, 2015; Kaplan et al., 2023).

Among many aspects, significant research has been devoted to exploring the impact of personal factors on trust in automation. These studies have revealed that human-related factors, such as cultural values, attentional control, mood, personality, propensity to trust, and expectations toward autonomous systems, significantly influence trust (Cai et al., 2022; Merritt et al., 2013; Sharan & Romano, 2020; Stokes et al., 2010; Zhou et al., 2020). This suggests that an individual's trust may vary based on personal characteristics, even when interacting with the same automation system under identical conditions. Insights from these studies have steered our focus toward predicting individuals' trust in automation using personal characteristics.

Also, an expanding body of research is examining the dynamic nature of trust, acknowledging that an individual's trust can evolve throughout their interaction with autonomous technologies. Recent studies have identified distinct trust dynamics exhibited by different individuals. For instance, McMahon et al. (2020) identified two clusters, followers and preservers, based on trust-dependent behaviors, with followers showing a greater inclination to trust and follow group displayed higher initial trust, slower trust erosion after violations, and quicker trust recovery. Guo et al. (2020) and Guo and Yang (2021) recognized three trust dynamics patterns in episodic tasks: Bayesian decision-makers, oscillators, and disbelievers, each with unique trust adaptation. Bhat et al. (2023) found similar patterns in a sequential task setting and further suggested associations between personal characteristics and trust dynamics types.

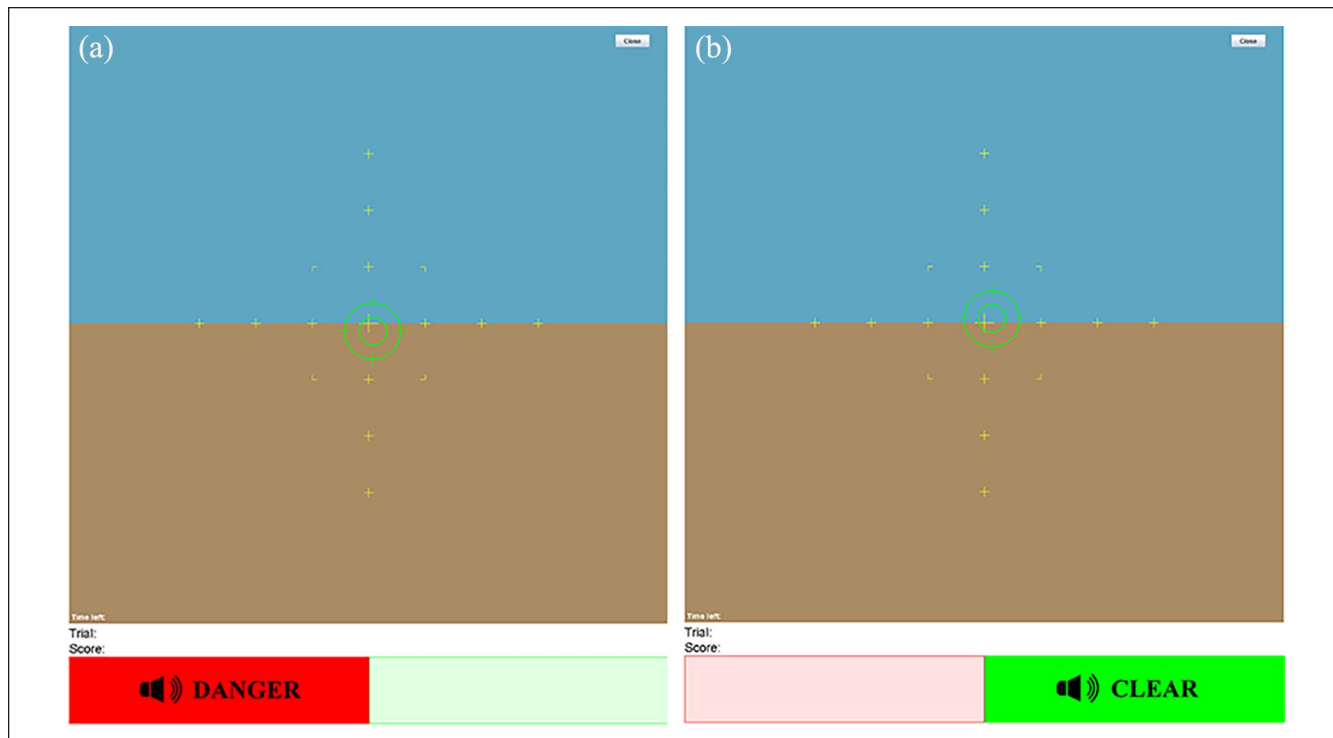
Despite extensive research on the effects of personal factors on trust and the identification of distinct trust dynamics clusters, developing a model to predict trust dynamics using personal factors has received limited attention. Given the

---

<sup>1</sup>University of Michigan, Ann Arbor, USA

### Corresponding Author:

Hyesun Chung, Industrial and Operations Engineering, University of Michigan, 1205 Beal Avenue, Ann Arbor, MI 48109-1382, USA.  
Email: hyesunc@umich.edu



**Figure 1.** Alerts from an automated threat detector: (a) detector giving “DANGER” alert and (b) detector giving “CLEAR” alert.

substantial empirical evidence that personal factors are significant antecedents of trust, this represents a notable research gap. To bridge this gap, building upon the study by Chung and Yang (2024), we aim to see how the key personal characteristics identified in the previous research could work to predict the type of trust dynamics that the user would exhibit. By using the data collected from the previous study (Chung & Yang, 2024), we have tested multiple different classification models to see how personal characteristics can be used to predict the trust dynamics type. After some experiments, we have concluded that the random forest classifier outperforms in performance and explainability. In sum, the primary aim of this study is to develop a predictive model using key personal characteristics.

## Methods

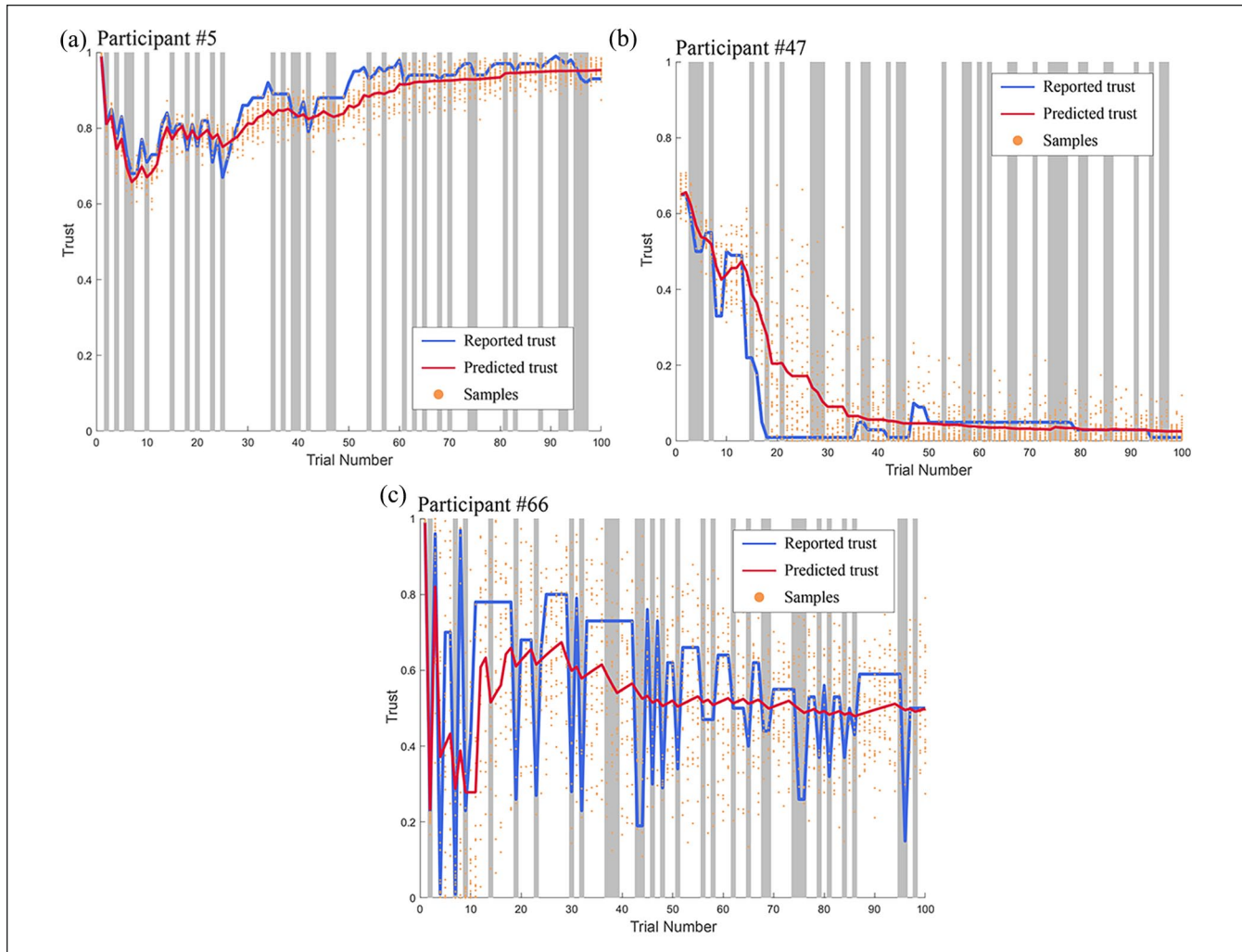
### Data used for analysis

In this study, we referred to the data collected from Chung and Yang (2024). In this previous work, a human-subject experiment involving 130 participants was conducted, where the participants had to perform a simulated surveillance task aided by an imperfect threat detector. They had to engage in recommendations, in contrast to preservers who showed lower levels of trust. Liu et al. (2021) differentiated skeptical from confident participants, noting that the confident 1To be consistent with early literature, we use the two terms automation and autonomy interchangeably in this paper while

acknowledging the difference between the two (Endsley, 2017). Two different tasks simultaneously: the tracking task and the detection task. The threat detector assisted the participants with the detection task by producing alarms (Figure 1). If it identified a threat, it provided a red display with the sound “Danger”; otherwise, if it did not identify any threat, the alert signal was green with the sound “Clear.” Each participant performed a total of 100 trials in an experiment. The reliability of the automated threat detector, which indicates the probability of it producing the correct alarm, ranged between 62% and 70%.

The authors collected a wide range of personal characteristics and trust dynamics data, which were measured during the experiment for a total of 100 times. Using the trust dynamics data, the authors identified three types of trust dynamics: Bayesian decision-makers, disbelievers, and oscillators. The Bayesian decision-makers are characterized as updating their trust in a Bayesian manner and showing a relatively higher level of trust throughout the 100 trials (Figure 2a). Conversely, the disbeliever group is marked by consistently exhibiting lower trust in the automated threat detector (Figure 2b). Lastly, the oscillator group is named as such because they show highly fluctuating and oscillating changes in their trust ratings, making it challenging to accurately predict their real-time trust level (Figure 2c).

Subsequently, they conducted ANOVAs (Analyses of Variance) to determine the differences in personal characteristics across the clusters. As a result, they identified seven dimensions of personal characteristics that show significant



**Figure 2.** Three distinct clusters of trust dynamics: (a) BDM, (b) disbeliever, and (c) oscillator.  
 Note. The figures are copied from Chung and Yang (2024).

differences across the three cluster groups: masculinity (from the measures of culture), positive affect (from the measures of mood), extraversion, neuroticism, and intellect (from the measures of personality), performance expectancy (from the measures of expectancy toward autonomous systems), and high expectations (from the measures of the perfect automation schema). The oscillator group showed the highest ratings for masculinity, positive affect, and extraversion. On the other hand, the disbeliever group had the highest average rating for neuroticism, but the lowest ratings for intellect, performance expectancy, and high expectations. For more details of the study protocol and the results, please refer to Chung and Yang (2024).

We used a subset of the data collected from 130 participants, specifically ratings for the seven personal characteristics (Table 1), which will be used as input variables in developing a prediction model. For the target variable, we used the cluster group to which each participant was assigned.

There are three classes for this target: Bayesian decision-maker (BDM), disbeliever, and oscillator.

### Prediction model

Using the data collected from the aforementioned study, we have conducted experiments to validate the key personal characteristics that have been found to be significantly different across the three clusters. Our goal was to test whether these can be good predictors of the type of trust dynamics. To accomplish this, we tested the data using multiple existing classification algorithms: Random Forest, multinomial logistic regression, Support Vector Machine (SVM), XGBoost, and Naive Bayes.

Before fitting the data, we first scaled the input variables using the standard scaling method. Following that, we divided the data into training and test sets in an 80:20 split. Since the cluster distribution is imbalanced, when

**Table 1.** List of Personal Characteristics Used as Predictor Features.

Feature	Definition	Reference
Masculinity	One dimension of cultural value that examines the extent to which one's cultural beliefs support traditional views of masculine and feminine traits	Cultural Values Scale (Yoo et al., 2011)
Positive affect	One dimension of mood that measures the degree of one's positive emotions, such as cheerfulness, enthusiasm, and energy	Positive and Negative Affect Schedule (Watson et al., 1988)
Extraversion	One dimension of personality that measures a trait of being outgoing and energetic	Mini-International Personality Item Pool (Donnellan et al., 2006)
Neuroticism	One dimension of personality that measures the tendency to experience negative emotions such as anxiety, anger, and frustration	Mini-International Personality Item Pool (Donnellan et al., 2006)
Intellect	One dimension of personality that measures the trait of having broad interests and being imaginative	Mini-International Personality Item Pool (Donnellan et al., 2006)
Performance expectancy	One dimension of expectancy toward autonomous systems that measures the degree to which one expects that utilizing the autonomous system will lead to improvements in work performance	Unified Theory of Acceptance and Use of Technology (Venkatesh et al., 2003)
High expectations	One dimension of the perfect automation schema that measures one's belief that the automation will perform with near-perfect reliability	Perfect Automation Schema (Merritt et al., 2015)

**Table 2.** Algorithm Performance.

Algorithm	Average accuracy (five-fold cross-validation)	Accuracy on test dataset
Random Forest	0.749	0.731
Multinomial Logistic Regression	0.760	0.731
SVM	0.750	0.731
XGBoost	0.673	0.692
Naïve Bayes	0.731	0.731

dividing the data, we performed stratified random sampling to maintain the class distribution of the original dataset in both the training and test datasets. Within the training set, we conducted hyperparameter tuning using a five-fold cross-validation method, aiming to identify hyperparameters that would maximize model performance. In all algorithms, the seven personal characteristics dimensions served as the input variables, and the output variable was the cluster type, which included a total of three classes. In the end, we comparatively evaluated the average accuracy of the five-fold cross-validation, accuracy on the test dataset, and the confusion matrix that resulted from the model prediction.

## Results

The primary goal of this study was to develop a classification model that predicts the type of trust dynamics using seven key personal characteristics. We tested with five different classification algorithms, and Table 2 tabulates the cross-validation scores and prediction accuracy for the test dataset for each algorithm.

From the results (Table 2), we confirmed that personal characteristics features can be used to predict the type of trust dynamics. Among the four algorithms, considering both the average accuracy for the five-fold cross-validation and the accuracy on the test dataset, the Random Forest, multinomial logistic regression, and SVM models seem to outperform the other two. Subsequently, we comparatively evaluated the confusion matrices of the different models. This evaluation was crucial because, given the imbalanced distribution of the dataset, being able to accurately predict the minority classes is an important criterion when choosing the final model. Figure 3a to c present the confusion matrices of the Random Forest, multinomial logistic regression, and SVM model, respectively.

As we compared the confusion matrices of the Random Forest Classifier, multinomial logistic regression, and SVM, we found that the Random Forest Classifier outperforms the others in terms of predicting the disbelievers and oscillators. Among 25 actual disbelievers, it correctly identified 20; among 14 actual oscillators, it correctly predicted 12. The results of the multinomial logistic regression (Figure 3b) and SVM (Figure 3c) show that both of the classifiers predominantly predict the

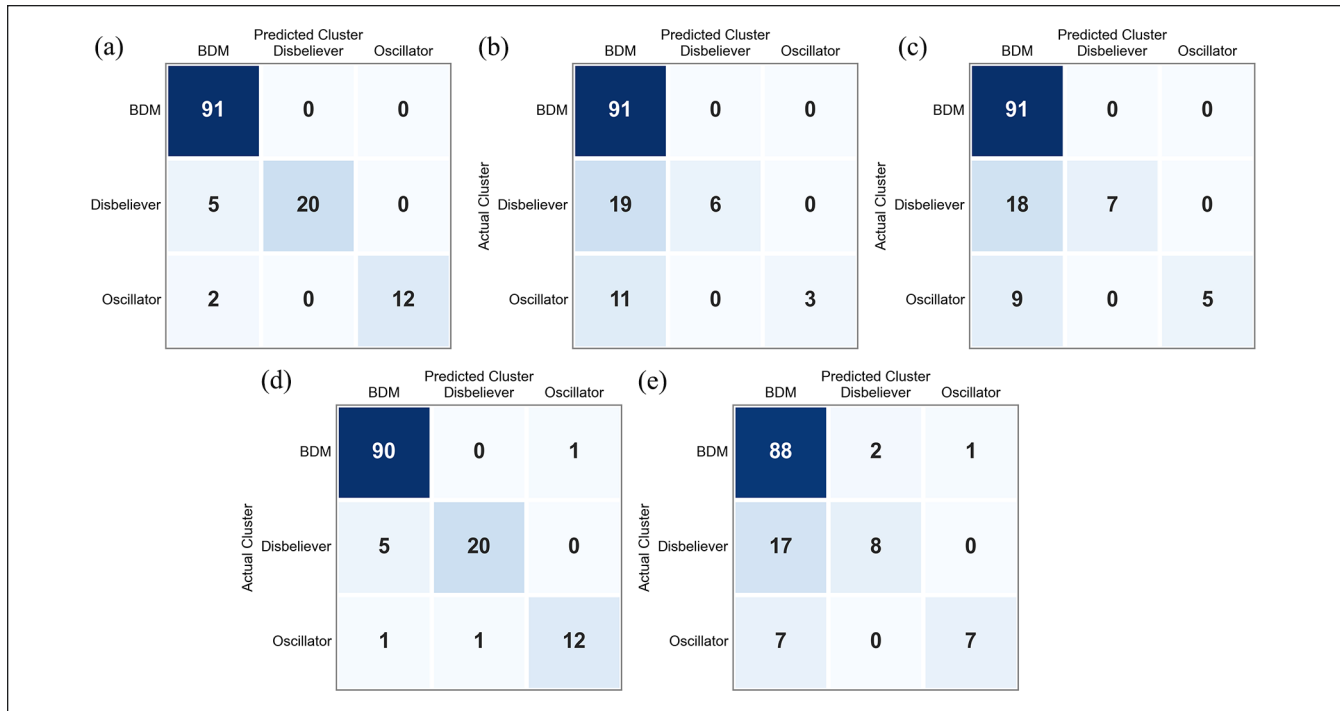


Figure 3. Confusion matrix: (a) Random Forest, (b) Multinomial Logistic Regression, (c) SVM, (d) XGBoost, and (e) Naive Bayes.

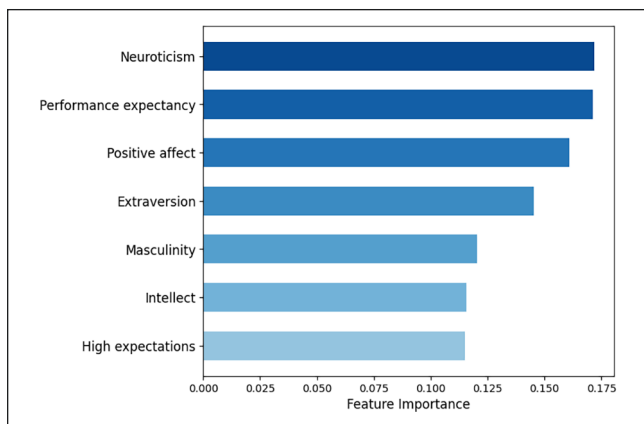


Figure 4. Feature importance scores.

majority class, the BDM group. Consequently, we conclude that, given the dataset, the Random Forest Classifier is the most appropriate classifier for predicting the trust dynamics type using personal characteristics. Figure 4 shows the feature importance scores in descending order.

### Discussion

This study aimed to develop a classification model capable of predicting the type of trust dynamics a user would exhibit based on personal characteristics. Building upon the data and results derived from a previous study (Chung & Yang, 2024),

we utilized seven personal characteristics (masculinity, positive affect, extraversion, neuroticism, intellect, performance expectancy, and high expectations) to develop the model. Subsequently, we comparatively evaluated the classification performance of different models.

Our results demonstrated that, considering overall accuracy and the ability to discern minority groups (i.e., disbelievers and oscillators), the Random Forest Classifier would be the most appropriate approach. Additionally, this classifier is superior in terms of explainability, as it allows for a deeper understanding of which features play more significant roles in classifying users (Figure 4). In this study, neuroticism and performance expectancy, the dimensions in which disbelievers scored lower, recorded the highest feature importance scores. Following these two, positive affect and extraversion, the personal characteristics dimensions in which oscillators scored higher, also had high feature importance scores.

The performance levels observed in this study suggest the potential to predict the type of trust dynamics an individual may exhibit by administering a set of questions about personal characteristics. This indicates the possibility of developing personalized trust prediction algorithms for each trust dynamics type, which could ultimately facilitate the creation of trust-aware agents. In essence, our classification model can initially identify individuals likely to exhibit specific trust patterns, allowing for adjustments in automation to accommodate their unique characteristics.

To elaborate, depending on the prediction results, system developers could infer users' initial trust levels and how they might adjust their trust levels as they interact with automated systems. Subsequently, based on these predictions, appropriate system designs that support trust calibration could be implemented. For instance, if a person is expected to be a disbeliever, yet the system is demonstrably reliable, the agent could provide additional prompts to encourage more active use of the automation, thereby preventing disuse. Conversely, if a user is expected to be a Bayesian decision-maker, yet the system is prone to unreliability, extra measures should be taken to prevent overtrust in the automation.

Moreover, the results from this study could be utilized to enhance the algorithm for predicting temporal trust in real-time. Although Guo and Yang (2021) and Guo et al. (2020)'s Beta random prediction model has proven effective in many cases, the oscillator group has been identified in multiple studies (Bhat et al., 2023; Chung & Yang, 2024; Guo et al., 2020; Guo & Yang, 2021). This group is characterized by a group of people, whose trust is challenging to predict due to their fluctuating patterns. Now that the classification model proposed in this study provides a method to predict this group of people beforehand, the trust prediction algorithm for this particular group can be improved by incorporating personal characteristics.

The study is open to further improvements. Additional efforts in model tuning and enhancement using ensemble methods could improve overall model accuracy. In addition, if these classification models are specifically aimed at identifying certain groups, such as oscillators, more work should be done to avoid overlooking these groups. The Balanced Random Forest Classifier could be a suitable alternative in this case. Unlike the classic Random Forest Classifier, it draws a bootstrap sample from the minority class and samples an equivalent number from the majority class with replacement. This approach may not necessarily yield high accuracy since it may focus on classifying minority classes and could misclassify many BDMs as other groups. Nonetheless, it could be highly successful in recalling all minority classes (i.e., disbelievers and oscillators). Consequently, if missing a minority group is a significant concern, the balanced classifier would be preferable to the classic one.

Overall, our study offers comprehensive insights into the diverse trust dynamics exhibited by different individuals, emphasizing the importance of specific personal characteristics as significant predictors of trust dynamics.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This material is based upon work supported by National Science Foundation under Grant No. 2045009.

### References

- Bhat, S., Lyons, J. B., Shi, C., & Yang, X. J. (2023). *Evaluating the impact of personalized value alignment in human-robot interaction: Insights into trust and team performance outcomes* [Conference session]. Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction.
- Cai, W., Jin, Y., & Chen, L. (2022). *Impacts of personal characteristics on user trust in conversational recommender systems* [Conference session]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.
- Chung, H., & Yang, X. J. (2024). *Associations between trust dynamics and personal characteristics* [Conference session]. 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS).
- Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipp scales: Tiny-yet-effective measures of the big five factors of personality. *Psychological Assessment, 18*(2), 192–203.
- Endsley, M. R. (2017). From here to autonomy: Lessons learned from human-automation research. *Human Factors, 59*(1), 5–27. <https://doi.org/10.1177/0018720816681350>
- Guo, Y., & Yang, X. J. (2021). Modeling and predicting trust dynamics in human-robot teaming: A bayesian inference approach. *International Journal of Social Robotics, 13*(8), 1899–1909.
- Guo, Y., Zhang, C., & Yang, X. J. (2020). *Modeling trust dynamics in human-robot teaming: A Bayesian inference approach* [Conference session]. Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527.
- Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors, 63*(7), 1196–1229.
- Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434.
- Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors, 65*(2), 337–359.
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.
- Liu, J., Akash, K., Misu, T., & Wu, X. (2021). *Clustering human trust dynamics for customized real-time prediction* [Conference session]. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC).
- McMahon, G., Akash, K., Reid, T., & Jain, N. (2020). On modeling human trust in automation: Identifying distinct dynamics through clustering of markovian models. *IFAC-PapersOnLine, 53*, 356–363.
- Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534.
- Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. *Human Factors, 57*(5), 740–753.

- 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.
- Stokes, C. K., Lyons, J. B., Littlejohn, K., Natarian, J., Case, E., & Speranza, N. (2010). *Accounting for the human in cyberspace: Effects of mood on trust in automation* [Conference session]. 2010 International Symposium on Collaborative Technologies and Systems.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27, 425–478.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070.
- Yoo, B., Donthu, N., & Lenartowicz, T. (2011). Measuring hofstede's five dimensions of cultural values at the individual level: Development and validation of cvscafe. *Journal of International Consumer Marketing*, 23(3–4), 193–210.
- Zhou, J., Luo, S., & Chen, F. (2020). Effects of personality traits on user trust in human–machine collaborations. *Journal on Multimodal User Interfaces*, 14, 387–400.