A Human-Machine Trust Model Integrating Machine Estimated Performance

1st Shaojun Chen
Department of Automation
University of Science and Technology of China
Hefei, China
shaojunchen@mail.ustc.edu.cn

2nd Yun-Bo Zhao*
Department of Automation
University of Science and Technology of China
Institute of Artificial Intelligence
Hefei Comprehensive National Science Center
Hefei, China
ybzhao@ustc.edu.cn

3rd Yang Wang
Department of Automation
University of Science and Technology of China
Hefei, China
yang_wang@mail.ustc.edu.cn

4th Junsen Lu
Department of Automation
University of Science and Technology of China
Hefei, China
lujunsen@mail.ustc.edu.cn

2023 6th International Symposium on Autonomous Systems (ISAS) | 979-8-3503-1615-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISAS59543.2023.10164500

***Abstract*—The prediction of human trust in machines within decision-aid systems is crucial for improving system performance. However, previous studies have only measured machine performance based on its decision history, failing to account for the machine's current decision state. This delay in evaluating machine performance can result in biased trust predictions, making it challenging to enhance the overall performance of the human-machine system. To address this issue, this paper proposes incorporating machine estimated performance scores into a human-machine trust prediction model to improve trust prediction accuracy and system performance. We also provide an explanation for how this model can enhance system performance.**

To estimate the accuracy of the machine's current decision, we employ the KNN(K-Nearest Neighbors) method and obtain a corresponding performance score. Next, we report the estimated score to humans through the human-machine interaction interface and obtain human trust via trust self-reporting. Finally, we fit the trust prediction model parameters using data and evaluate the model's efficacy through simulation on a public dataset. Our ablation experiments show that the model reduces trust prediction bias by 3.6% and significantly enhances the overall accuracy of human-machine decision-making.

***Index Terms*—human machine trust, machine learning, decision-aid systems, KNN**

I. Introduction

Automated technology has found widespread and deep application in both industrial and civilian domains [1], [2]. The implementation of automated technology is achieved through machines, and with the application of AI technology in machines, they have gained decision-making capabilities. However, these decisions are uncertain and not entirely reliable. Thus, autonomous decision-making systems still require human supervision and coordination to ensure optimal system performance [3]. Human-machine decision-aid systems are a prime example of such systems. The main workflow is shown in Figure 1, where the machine presents decision results to humans using machine learning methods, and humans make judgments and decide whether to adopt the machine's decision In human-machine collaborative decision-aid systems, trust

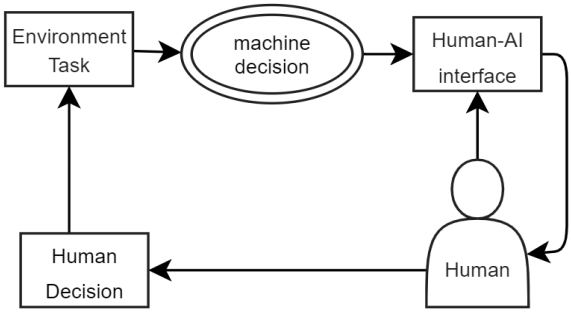

Fig. 1. The workflow of human-machine decision-aid systems

in machine decision-making is critical for fostering effective interaction and cooperation between humans and machines [4]. Over-reliance on machine decisions can result in catastrophic consequences when the machine makes an error, while a lack of trust in machine decision-making undermines the benefits of automated decision-making, leading to reduced efficiency of the human-machine system as a whole [4]–[6]. Hence, guiding trust is of considerable research significance in human-machine collaborative decision-aid systems [7], [8].

Guiding trust in human-machine collaborative decision-making systems requires a predictable, reliable, and quantitative human-machine trust model [9]. A human-machine trust model typically quantitatively describes the relationship between human trust in the machine and the machine's performance. A differential equation trust model was proposed in [10], which uses the machine's accuracy or error rate, including false positive and false negative rates, as a measure of machine performance. The model also accounts for accumulated trust, gender, and cultural factors to construct the human-machine trust model. However, since the machine's historical decision accuracy remains constant or fluctuates slightly over

This work was supported by the National Key Research and Development Program of China (No. 2018AAA0100801).

* Corresponding author

time, this model fails to capture dynamic changes in machine performance.

A trust model based on POMDP(Partially Observable Markov Decision Process) was proposed in [11], which treats trust as a hidden state variable, uses the machine's previous decision correctness as a measure of machine performance, and considers transparency and other factors to construct the trust model. This model can fully capture dynamic changes in the machine's historical decision accuracy. However, the model only considers the correctness of the previous decision and lacks a quantitative characterization of the relationship between the current decision and the previous decision. Additionally, the parameterized algorithm used in this model is highly sensitive to initial values, leading to significant fluctuations in trust prediction.

Overall, research on trust models in human-machine collaborative decision-making systems is insufficient, particularly in the existing literature, where machine performance descriptions are limited to historical machine decision-making, such as accuracy and previous decision correctness, without characterizing the current machine decision-making state. As a result, human perception of machine performance is delayed, leading to a bias that can significantly reduce the performance of the human-machine system in some cases. Therefore, it is necessary to incorporate the current machine decision-making state into the trust model to reduce the bias in trust prediction and improve the overall performance of the human-machine system.

Inspired by a quantitative measurement method for classifier credibility in machine learning proposed in [12], we utilized the machine's decision history as a sample library and compared the current decision result with the decision history to obtain an estimated performance score of the machine's current decision as a characterization of the machine's decision-making state. We improved the method proposed in [12] and integrated this score into the human-machine trust prediction model to predict human trust. Experimental results demonstrate that our proposed trust model, which incorporates machine estimated performance , can effectively predict human trust compared to other trust models, while also improving the overall performance of the human-machine system.

II. Human Machine Trust Model

A. Machine Estimated Performance Score

Decision-aid systems involve the selection of the most appropriate option from a set of existing options using machine learning techniques. As such, they can be viewed as a classification problem in machine learning. To this end, we organize the machine's decision history as follows: firstly, for all decision histories $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, where $x_i \in X$ represents the sample and $y_i \in Y$ represents the correct classification of the sample, we partition all samples based on their classifications, defining the set $H_l := xi|1 \leq i \leq n, y_i = l$ for each $l \in Y$. Then, for each $l \in Y$, we organize the elements of H_l into a KD tree T_l.

We calculate the estimated performance score of the machine's current decision as follows: given the current task x_N and the machine's decision result y_N, we first place x_N into all KD trees in the decision history. In each KD tree T_l, we calculate the distance between X and the nearest element, d_l. In reference [12], the method for calculating the credibility of machine decisions T_0 is as follows:

$$T_0 = \frac{min\{d_l, l \in Y, l \neq y_N\}}{d_{y_N}} \tag{1}$$

where d_{y_N} denotes the distance between the current sample and the nearest element in the KD tree corresponding to the machine's decision result y_N.

Under the calculation method in [12], regardless of the machine's decision result, this measurement method always yields a positive number. When this number is greater than 1, it means that the probability of the machine's decision being correct is higher, and when it is less than 1, the machine's decision is more likely to be incorrect. However, in a decision-aid system, it is crucial to pay more attention to the possible errors made by the machine. Therefore, we propose an improved method that considers the consequences of both correct and incorrect machine decisions. Additionally, when the machine makes an incorrect decision, we prefer the estimated performance score to be negative, as this can better incorporate the characterization of the machine's decision-making state into the trust model.Taking the above two points into consideration, we have made some improvements to the method proposed in [12]. First, we define the set Q as follows:

$$Q := \{d_l | d_l < d_{y_N}, l \in Y\}. \tag{2}$$

The estimated performance score of machine decision is defined as follows:

$$T = \begin{cases} \Sigma_{d \in Q} - card(Q) \cdot d_{y_N}, & if\ card(Q) \neq 0 \\ T_0 \cdot d_{y_N} - d_{y_N}, & if\ card(Q) = 0 \end{cases} \tag{3}$$

where $card(Q)$ denotes the cardinal number of set Q

Remark: The underlying principle of the above method is that the smaller the value of d_{y_N}, the closer the machine's current decision result is to the correct historical decision results, and the better the estimated performance score. Other categories' nearest distances are also included to make the relative value more reasonable. When Q is empty, the calculation method in this paper is not fundamentally different from that in literature [12]. However, when Q is not empty, especially when there are multiple elements in Q, this indicates that there are multiple classification results from historical decision data that are better than the current decision result. The method in literature [12] only uses one of them. The improved method proposed in this paper utilizes all classification results and strengthens the score of machine decision-making errors, making human judgment more vigilant and improving the overall performance of the human-machine system.

B. Human Machine Trust Model Equation

Machine performance is the most crucial factor that affects human-machine trust. Therefore, the trust prediction model proposed in this paper characterizes the relationship between trust and machine performance without considering cultural, gender, or other influences. Jonker and Treur [13] suggested that the change in trust is directly proportional to the difference between experience and past trust. Here, historical experience refers to the machine's past decision performance, which has been adequately addressed in relevant research literature [10], [11]. Building on this, we integrate the machine's current estimated performance to formulate a new trust prediction model, as follows:

$$\begin{aligned} S(n+1) - S(n) = \alpha_1(R(n) - S(n)) \\ + \alpha_2(T(n) - S(n)) \\ + \alpha_3 C(n) + \alpha_4 D(n) + \alpha_5 S(n) \end{aligned} \tag{4}$$

where S represents the trust value that humans have in machine decision-making, ranging from 0 to 100. R represents the accuracy of the machine's historical decisions, while C and D indicate whether the machine's decision at the previous moment was correct(1)or incorrect(0). The subscript n refers to the previous moment, and $n+1$ to the next moment. The remaining variables are coefficients that require estimation in the model.

Remark: The proposed model, MEPTM (Machine Estimated Performance Trust Model) considers the influence of historical decision accuracy, which is in line with previous literature. Moreover, we provide a more detailed classification for the correctness of the machine's decision at the previous moment: generally, if the machine's decision is correct, there will be a slight increase in the human's trust in the machine, while an incorrect decision will result in a more significant decrease in the trust value due to a greater negative response to the machine's wrong decision. Hence, we have developed different coefficients to address this situation. Additionally, we have accounted for the inertia of human decision-making by introducing the variable D into the model.

When estimating performance, we measure the distance between the machine's decision and the correct decision, rather than using division as a performance score, which is the approach used in literature [12]. Our trust model justifies this change. When the machine makes an incorrect decision, the human's trust in the machine should decrease. Using a ratio to calculate the estimated performance score could result in a positive score greater than the current trust value, which is not reasonable, especially if the human's initial trust value is low. However, our proposed method results in a negative estimated performance score that will decrease the trust value when incorporated into the trust model, making this calculation method more reasonable.

C. Parameterizing the trust model

Formula (4) involves five parameters that must be estimated by fitting the data. A simple analysis suggests that these factors can be treated as independent variables, indicating that this is a linear model. To obtain the parameters of the trust model, we first employ the trust self-reporting method to obtain the trust values that humans have in machine decision-making for each trial. Next, we use the data fitting command in the open-source Python library scikit-learn (sklearn) to estimate the parameters of the trust model.

III. Experimental validation

To validate the effectiveness of the trust model integrated with machine performance estimation, we conducted real experiments with volunteer participants. The experiments were divided into two groups: one group received the machine's estimated performance score through the interface, while the other group did not receive any performance score and served as the control group. Each group was tested 70 times. The entire experiment process is shown in Figure 2:

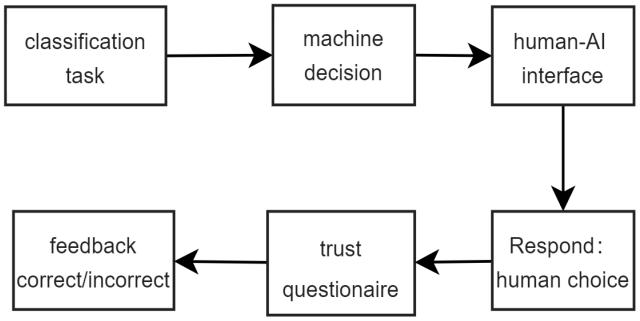

Fig. 2. Experiment process of MEPTM

The experiment was comprised of four main parts: First, the machine made a decision for the current classification task and reported the result to the human through a human-machine interaction interface. Secondly, the human evaluated the machine's decision outcome and made a final choice. Then, the human completed a trust questionnaire to report their level of trust in the machine's decision for that particular task. Finally, the system provided feedback to the human on the correct category of the current classification task.

A. Experiment settings

- The classification task is selected from the public dataset "Digits" in scikit-learn, which is a classic dataset for recognizing handwritten numbers and contains a total of 1797 images [14]
- We utilized the Naive Bayes algorithm to make decisions on behalf of the machine. During the training phase, we used the first 200 data samples from the dataset, which led to the accuracy rate of the Naive Bayes algorithm fluctuating between 65% and 85%
- At the same time, we also obtained the previous 200 samples and their correct classifications, which were put into the algorithm for estimating the machine's performance to initialize the training of the performance estimation algorithm.

- In the experiment, we excluded the pre-trained 200 samples and randomly selected 70 samples from the remaining dataset for testing.
- Prior to the experiment, volunteers were provided with information regarding the type of machine decision-making algorithm, the accuracy of machine decision-making, the significance of the machine's performance estimation, and the goal of the human-machine system to make correct decisions in the shortest amount of time.

B. Results

In the experiment, we utilized the outcomes of the initial 50 trials as fitting data to parameterize the trust prediction model, while the outcomes of the final 20 trials were utilized to validate the model.

Firstly, we directly predicted the trust values. The results of the experimental and control groups are shown in Figures 3 and 4, respectively. The trust prediction model for the control group is as follows:

$$S(n+1) - S(n) = \alpha_1(R(n) - S(n)) + \alpha_3 C(n) + \alpha_4 D(n) + \alpha_5 S(n). \tag{5}$$

Compared to the MEPTM model, the control group's model lacks the machine performance prediction score.

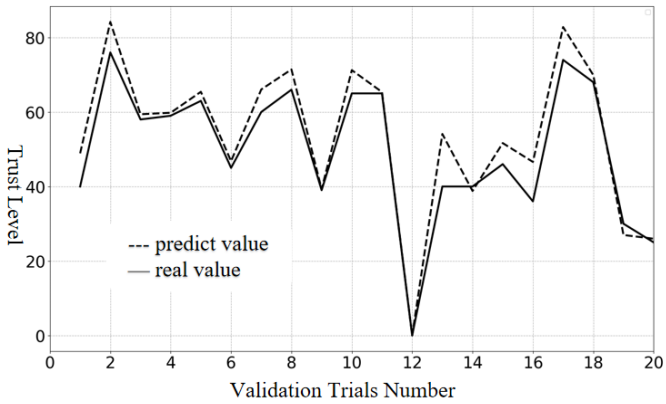

Fig. 3. Trust prediction of MEPTM

The results from Figure3 and Figure4 indicate that the prediction deviation of MEPTM is more stable, fluctuating closely around the true values, while the control group demonstrates a smaller or larger difference from actual values in some cases. The variance between actual and true values was calculated for the 20 experimental verifications, yielding variances of 1.32 and 1.37 for the experimental and control groups, respectively. This leads us to conclude that MEPTM is more robust in predicting trust values.

Furthermore, we observed an interesting phenomenon where the range of actual trust levels in the control group is smaller than that in MEPTM, fluctuating around 60. This suggests that in MEPTM, humans have more confidence in their own decision-making. When the machine makes a wrong decision, human trust values will significantly decrease, while a correct

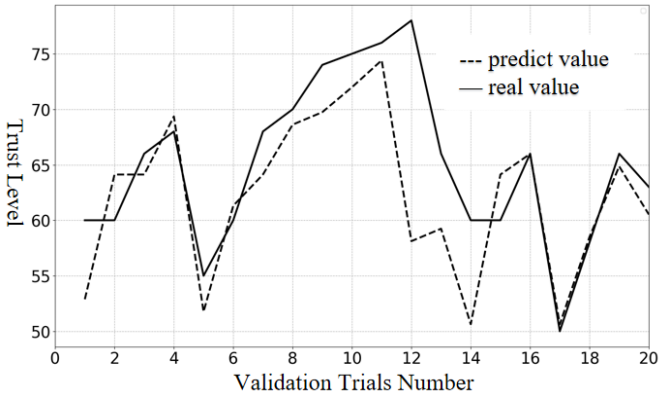

Fig. 4. Trust prediction of control group

decision by the machine will lead to a significant increase in human trust values.

In a decision support system, human decision-making is the ultimate decision of the human-machine system. It is worth noting that the significance of human-machine trust lies in its ability to impact human decision-making, which consequently affects the accuracy of the decision support system. Additionally, during the practical use of human-machine support systems, it is unattainable to complete a trust questionnaire for every task, thereby hindering the acquisition of the actual trust values.

Therefore, considering the practical use of decision support systems and further highlighting the advantages of MEPTM trust prediction, we obtain predicted trust values through the trust model and use them to calculate human action selection. We use the trust-action model proposed in [15] to accomplish this, which is primarily a logistic regression model that maps trust values and accuracy information to the probability of discrete actions. An action is taken when the probability is greater than 0.5, and another action is taken when the probability is less than 0.5. In this paper, we represent humans accepting machine decisions with 1 and rejecting them with 0. The results of the two experiments are shown in Figures 5 and 6.

From Figures 5 and 6, it can be seen that out of 20 experiments, MEPTM correctly predicted human decisions 19 times, with only one error where the probabilities of 0 and 1 were still within the same order of magnitude. In contrast, the control group only made 16 correct predictions. MEPTM's prediction error rate was 5%, while the control group's error rate was 20%, a difference of four times. On one hand, without machine estimated performance , it becomes difficult to predict the impact of trust on human actions, which become more random. On the other hand, MEPTM strengthens the influence of trust on human decisions through machine performance estimation scores, resulting in a significant reduction in the error rate of human decisions. Therefore, solely looking at the variance values, the difference between MEPTM and the control group is not significant, with an improvement of only about 3.6%. However, these prediction deviations can have a

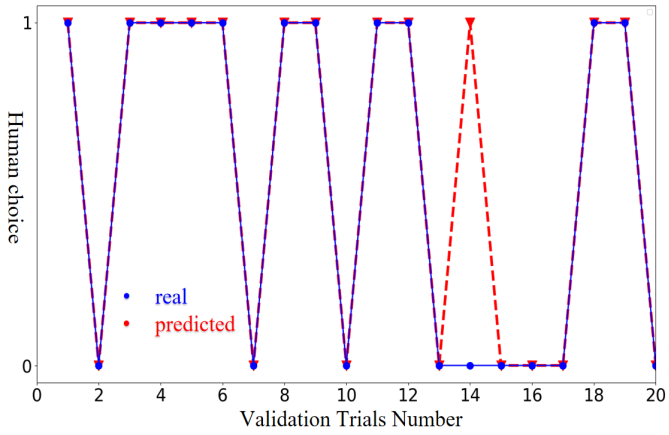

Fig. 5. The human choice prediction of MEPTM

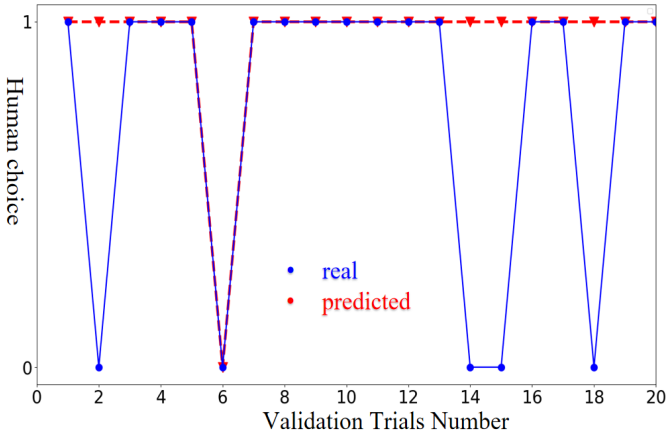

Fig. 6. The human choice prediction of control group

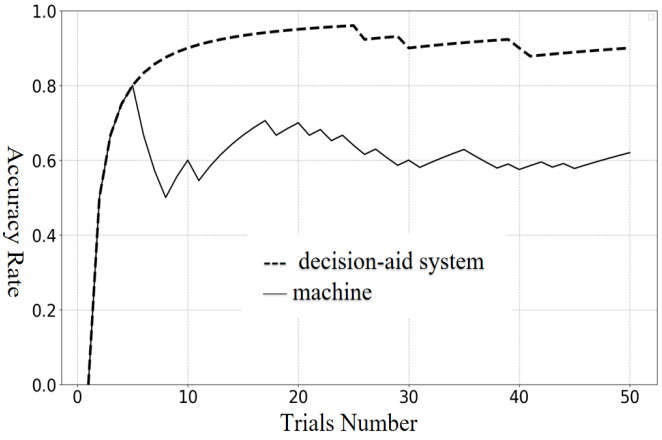

Fig. 7. The human choice prediction of MEPTM

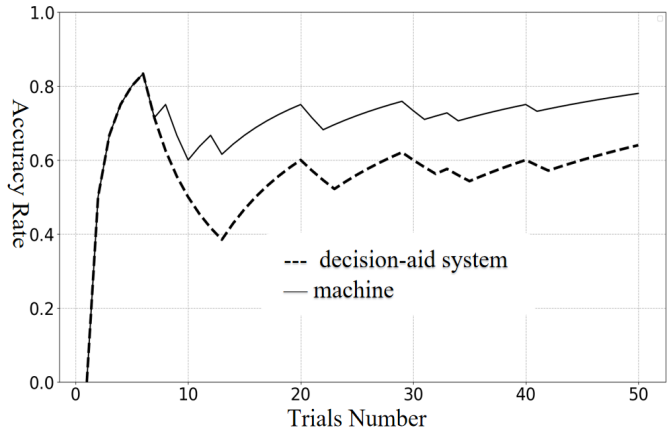

Fig. 8. The human choice prediction of control group

significant impact on human decision-making.

In addition to the precision of trust prediction by the MEPTM model, we conducted a further investigation into the holistic performance of the decision-aid system subsequent to the incorporation of machine performance estimation scores. We analyzed the initial 50 experiments in both the MEPTM and control groups, employing the accuracy of the decision-aid system as a criterion for assessing system performance. The experimental outcomes for the MEPTM and control group are depicted in Figures 7 and 8, respectively.

As the classification tasks performed by the control and experimental groups were randomly selected and therefore not identical, our evaluation criterion was the difference between the overall performance of the human-machine system and the performance of the machine alone. In the case of MEPTM, the overall performance of the human-machine system exceeded that of the machine alone, whereas in the control group, the opposite was true. It is evident that MEPTM can significantly enhance the overall performance of the human-machine system. Furthermore, an interesting conclusion can be drawn: the performance of the decision-aid support system is not necessarily better than that of the machine alone.

Next, we will elucidate why the human-machine performance of MEPTM outperforms that of the control group. Specifically, in the initial 25 experiments of the control group, we depicted the machine decision curve (with 1 indicating correct machine decisions and 0 indicating incorrect ones), the human execution curve (with 1 denoting execution of machine decisions by humans and 0 representing rejection of machine decisions), and the decision curve of the human-machine system (with 1 denoting correct decisions and 0 denoting incorrect ones), as illustrated in Figure 9.

It can be observed that when the machine's decision results change, humans tend to adjust their own decisions accordingly. For instance, when the machine shifts from making correct decisions to making incorrect decisions, humans tend to reject the machine's decision in the next task. Conversely, when the machine changes from making incorrect decisions to making correct decisions, humans tend to execute the machine's decision in the next task. However, the issue here is that while the machine's decision results change only in the historical sense, human decisions change in the present moment, leading to a delayed effect. When the machine's decision results oscillate periodically (alternating between correct and incorrect), the

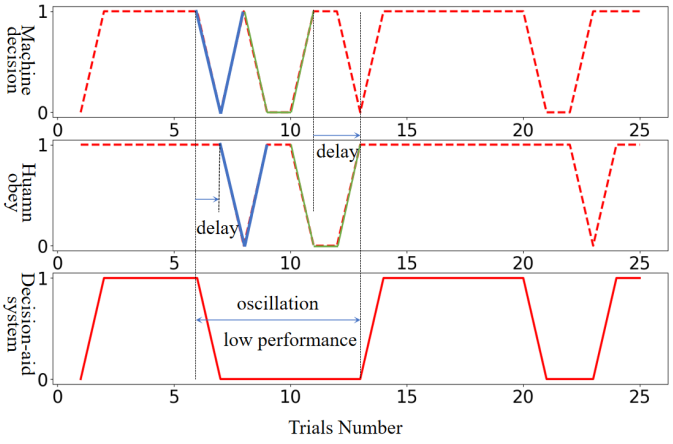

Fig. 9. The decision curve of control group

trust level of humans in the machine fluctuates between over-reliance and lack of trust, causing a significant decline in the overall human-machine performance. The root cause of this phenomenon is that in the control group, the lack of machine estimated performance results in minimal historical accuracy changes, making it difficult for humans to accurately evaluate the machine's current true ability. Therefore, humans rely heavily on the machine's previous decision results to make judgments, resulting in cognitive bias. As depicted in Figure 10, the period of low overall human-machine performance corresponds to the period of machine decision oscillation.

Likewise, we plotted the series of curves for the MEPTM model, as shown in Figure10.

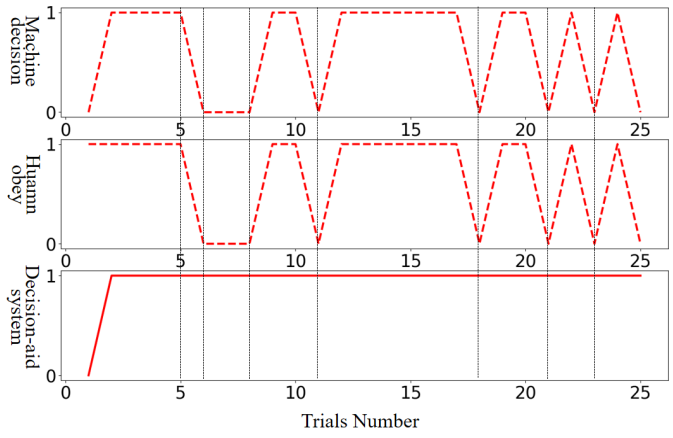

Fig. 10. The decision curve of MEPTM

In the MEPTM model, synchronization between human and machine decision results eliminates any delay. Hence, while the machine's decisions in MEPTM may also exhibit oscillations, humans can track them in real-time, resulting in a significant reduction of their cognitive bias due to the machine's performance estimation scores. Consequently, the trust level between humans and machines remains in a well-matched and appropriate state.

IV. Conclusion

In this paper, we present a novel approach to enhance trust prediction and improve the performance of decision-aid systems by incorporating the machine's estimated performance score into the trust model. Our key contribution lies in refining the calculation method for the estimated performance score and seamlessly integrating it into the human-machine system to form the MEPTM. We leverage trust self-reporting to parameterize the trust model with data and conduct experiments to validate the efficacy of the MEPTM while reducing bias in trust prediction.

References

[1] Y. Wang, K. N. Plataniotis, A. Mohammadi, L. Marcenaro, A. Asif, M. Hou, H. Leung, and M. Gavrilova, "Perspectives on the emerging field of autonomous systems and its theoretical foundations," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pp. 1–5, IEEE, 2021.

[2] T. Joo and D. Shin, "Formalizing human–machine interactions for adaptive automation in smart manufacturing," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 529–539, 2019.

[3] Y. Zhao, Y. Kang, and J. Zhu, *Autonomy Theory and Methods of Human-Machine Hybrid Intelligent Systems*. Science Press, 1st ed., 2021.

[4] J. Y. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 13–29, 2014.

[5] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[6] M. Hou, "Enabling trust in autonomous human-machine teaming," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pp. 1–1, IEEE, 2021.

[7] A. Kolling, P. Walker, N. Chakraborty, K. Sycara, and M. Lewis, "Human interaction with robot swarms: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 9–26, 2015.

[8] L. G. Dizaji and Y. Hu, "Building and measuring trust in human-machine systems," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, pp. 1–5, IEEE, 2021.

[9] B. Gebru, L. Zeleke, D. Blankson, M. Nabil, S. Nateghi, A. Homaifar, and E. Tunstel, "A review on human–machine trust evaluation: Human-centric and machine-centric perspectives," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 5, pp. 952–962, 2022.

[10] W.-L. Hu, K. Akash, T. Reid, and N. Jain, "Computational modeling of the dynamics of human trust during human–machine interactions," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 6, pp. 485–497, 2018.

[11] K. Akash, G. McMahon, T. Reid, and N. Jain, "Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions," *IEEE Control Systems Magazine*, vol. 40, no. 6, pp. 98–116, 2020.

[12] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," *Advances in neural information processing systems*, vol. 31, 2018.

[13] C. M. Jonker and J. Treur, "Formal analysis of models for the dynamics of trust based on experiences," in *Multi-Agent System Engineering: 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99 Valencia, Spain, June 30–July 2, 1999 Proceedings 9*, pp. 221–231, Springer, 1999.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[15] A. Xu and G. Dudek, "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pp. 221–228, 2015.