IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 October 2023, accepted 28 November 2023, date of publication 5 December 2023, date of current version 13 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3339548

RESEARCH ARTICLE

Dynamic Selection of Reliance Calibration Cues With AI Reliance Model

YOSUKE FUKUCHI[1] **AND SEIJI YAMADA**[1,2], **(Member, IEEE)**
[1]Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan
[2]Department of Advanced Studies, The Graduate University for Advanced Studies (SOKENDAI), Hayama, Kanagawa 240-0115, Japan

Corresponding author: Yosuke Fukuchi (fukuchi@nii.ac.jp)

This work was supported in part by Japan Science and Technology Agency (JST) Core Research for Evolutional Science and Technology (CREST), Japan, under Grant JPMJCR21D4.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the National Institute of Informatics Application No. R4- 6- 1.

ABSTRACT Understanding what an AI system can and cannot do is necessary for end-users to use the AI properly without being over- or under-reliant on it. Reliance calibration cues (RCCs) communicate an AI's capability to users, resulting in optimizing their reliance on it. Previous studies have typically focused on continuously presenting RCCs, and although providing an excessive amount of RCCs is sometimes problematic, limited consideration has been given to the question of how an AI can selectively provide RCCs. This paper proposes vPred-RC, an algorithm in which an AI decides whether to provide an RCC and which RCC to provide. It evaluates the influence of an RCC on user reliance with a cognitive model that predicts whether a human will assign a task to an AI agent with or without an RCC. We tested vPred-RC in a human-AI collaborative task called the collaborative CAPTCHA (CC) task. First, our reliance prediction model was trained on a dataset of human task assignments for the CC task and found to achieve 83.5% accuracy. We further evaluated vPred-RC's dynamic RCC selection in a user study. As a result, the RCCs selected by vPred-RC enabled participants to more accurately assign tasks to an AI when and only when the AI succeeded compared with randomly selected ones, suggesting that vPred-RC can successfully calibrate human reliance with a reduced number of RCCs. The selective presentation of RCCs has the potential to enhance the efficiency of collaboration between humans and AIs with fewer communication costs.

INDEX TERMS Reliance calibration, reliance prediction, reliance calibration cue, trust calibration, explainable AI, human-AI interaction, human-AI collaboration.

I. INTRODUCTION

Machine learning (ML) has been increasingly integrated into artificial agents and has improved their performance in complex tasks. However, the blackbox nature of ML models makes it difficult for end-users to understand the behavior of such agents [1], [2]. Particularly, a lack of understanding about what agents can and cannot do leads users to over- or under-rely on them [3], [4]. Over-reliance, in which a human overestimates the capability of an AI agent, can cause misuse and task failure [5]. Under-reliance is also problematic because it results in disuse, increases human workload, and degrades total collaboration performance.

Previous studies attempted to adjust human reliance by providing communication cues or information elements used by humans to assess an AI's capability [6], which we call reliance calibration cues (RCCs). A challenge facing reliance calibration with RCCs lies in the decision of whether to provide an RCC. In typical previous studies, all available RCCs are provided continuously. However, it has also been revealed that giving an excessive amount of information on an AI can have negative consequences [7], [8], [9]. There is a trade-off between successful calibration and reducing the communication cost, but it remains an open question of how an AI can selectively provide RCCs.

This paper proposes the *verbal Predictive Reliance Calibrator* (vPred-RC), an algorithm for selectively providing RCCs (Fig. 1). The main idea of vPred-RC is that it selects

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

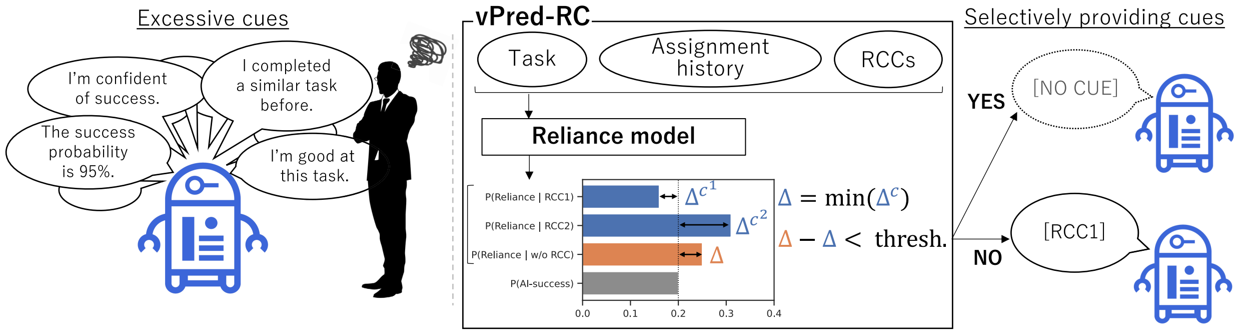

FIGURE 1. **Reliance calibration cues prevent human over/under-reliance on AI systems, but it can be problematic if the AI provides cues too often. *verbal* Predictive Reliance Calibrator (vPred-RC) enables AI system to selectively provide trust calibration cues. In vPred-RC, reliance model predicts probability that human will assign current task to AI. By considering both cases with and without cue provision, vPred-RC evaluates how much cue will contribute to reliance calibration and decides whether to provide it.**

whether to provide an RCC to avoid a wide gap between the human reliance rate and the AI's success probability. Here, the reliance rate is the probability that a human will assign a current task to the AI system. With a reliance model, which simulates user cognition of assigning a task to an AI or themselves, vPred-RC predicts reliance rates in both cases where an RCC is provided or not. By comparing the predicted reliance rates with the success probability (actual reliability), vPred-RC evaluates the impact of an RCC for reliance calibration. The reliance model is expected to enable an AI to provide information to users with an awareness of their potential needs for proper use of the AI.

This paper reports experiments for evaluating the reliance model and vPred-RC. First, to train the reliance model, we built a dataset of human decisions in a human-AI collaborative task in which RCCs were randomly provided. As a result of training, we found that the reliance model can predict human decisions with 83.5% accuracy. In the evaluation of vPred-RC, we focused on crowdworkers' decision accuracy, or how many times the workers assigned tasks that an AI could solve to the AI and did ones that the AI could not by themselves. The results show that the workers' accuracy was better with vPred-RC's selective RCCs than that of workers whose RCCs were randomly provided, suggesting that vPred-RC enables an AI to properly decide whether to provide RCCs by predicting and comparing reliance with and without an RCC.

The content of this paper is based on our previously published proceedings paper [10], in which we proposed Pred-RC. A main extension of vPred-RC is that it handles verbal cues as RCCs toward interactive human-robot collaboration, whereas Pred-RC considers a situation in which an AI system displays an indicator of its confidence rates as RCCs. In addition, this paper presents a more general formalization of the problem for vPred-RC to handle multiple RCC candidates. We newly conducted an experiment in a verbal and multiple-RCC setting, and the results show that vPred-RC can enable an AI system to properly decide whether to present an RCC and which RCC to present.

The remainder of this paper is structured as follows. Section II gives the background and reviews related work. Section III formalizes the problem of selectively providing RCCs and proposes vPred-RC. Section IV describes the preparation of a dataset for training the reliance model and reports an analysis of training results. Section V reports an evaluation of vPred-RC with the reliance model trained in section IV. Section VI provides discussions and limitations following the results of the evaluation. Section VII concludes this paper.

II. BACKGROUND

A. TRUST/RELIANCE CALIBRATION

Reliance is a concept relevant to trust, and it is sometimes studied inclusively. Trust is attitudinal and a psychological construct, while reliance focuses on the behaviors of humans, which is directly observable and thus an objective measure [11]. Although the main focus of this paper is reliance, this section reviews both trust and reliance calibration to highlight our research because of their close relevance.

There are various approaches to achieving trust/reliance calibration. Chen et al. focused on influencing human reliance by changing an agent's action. Their trust-POMDP is a computational model for deciding an action with awareness of human trust [12]. This paper considers the situation of calibrating reliance by explicitly communicating an AI's capability through RCCs. Previous studies have mainly focused on providing RCCs continuously [13], [14]. McGuirl and Sarter compared the effect of presenting dynamic system confidence with overall reliability only and found that the former can improve trust calibration [15].

Some studies revealed that providing an excessive amount of information on AI decision-making can have negative consequences. Ferguson et al. found that detailed explanations can cause information overload and lead to poorer trust than providing simple explanations [7]. Oh et al. noted that excessive explanation can harm the user experience [8]. Ehsan et al. discussed that whether an explanation that fits the context is more important than its length [9]. Okamura &

Yamada found that humans sometimes pay less attention to continuously displayed information. In their experiment, participants did not change their over-reliance in spite of a reliability indicator being continuously displayed to them. They also found that giving additional trigger signals was effective in resolving this problem [16]. Therefore, vPred-RC aims to calibrate reliance by not continuously but selectively providing RCCs.

Only a limited number of methods are proposed for selectively providing RCCs. A method proposed by Okamura & Yamada judges whether an AI should provide an RCC or not with "trust equations," logical formulae that mathematically express a human's over/under-reliance [16], [17]. A problem with this method is that its judgment depends only on how many times a human falsely assigns a task to an AI or him/herself, and it cannot capture the details of collaboration experiences such as in what task a human observed an AI's failure, when an AI provided RCCs, and what the tasks were. These experiences can affect human beliefs about an AI's capability. For example, an experience with an AI's success/failure on a task is more likely to influence human reliance in a similar task than a different task. In vPred-RC, a reliance model is trained to predict human reliance, taking into account the collaboration history between a human and an AI, and it is expected to capture these aspects.

B. RELIANCE ESTIMATION

A basic idea of vPred-RC is that inferring human reliance on an AI agent helps with the selective provision of RCCs. For example, an RCC that increases human reliance may be less effective if a human already has high reliance on an agent than if s/he has low reliance.

Muir's trust model refers to human intervention or takeover of a robot's action as an indicator of poor trust [18]. A user's decision on whether to assign a task to themselves or an AI is also used as an indicator of human trust [16], [17], and vPred-RC follows this approach.

Many methods have been proposed to estimate reliance/trust, but none of them can account for the impact of RCCs on human reliance, or the effects of RCCs that have already been shown to a human and how reliance changes if or unless an RCC is provided for a current task, which our reliance model aims to achieve.

III. SELECTIVELY PROVIDING TRUST CALIBRATION CUES

A. FORMALIZATION

This paper formalizes human-AI collaboration with selectively provided RCCs as a tuple $(x, \hat{\mathbf{c}}, c, d, y^*, y, p)$. Let us consider a situation in which a human sequentially performs a set of tasks $\{x_i\}_{i=1}^N$ with an AI agent, where i is the index of a task and N is the number of tasks. $\hat{\mathbf{c}}_\mathbf{i}$ is a set of potential RCCs for the AI system when x_i is given, and vPred-RC decides whether to provide $\hat{c} \in \hat{\mathbf{c}}_\mathbf{i}$. c_i is the RCC that the AI agent actually decides to provide to the human. $c_i = \hat{c}^{\mathrm{w/o}}$ represents that no RCC is provided.

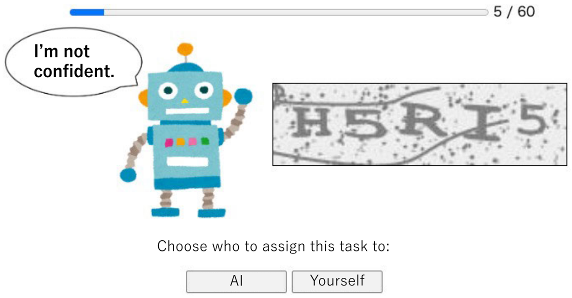

FIGURE 2. Screenshot of user interface for CC task.

The human observes (x_i, c_i, y_i^*) and decides whether to assign x_i to him/herself or the AI agent. Let $d_i \in \{\mathrm{AI}, \mathrm{human}\}$ be the agent to which a human decided to assign x_i. y_i^* is the desired result for x_i, and y_i is the actual result for x_i performed by d_i. $y_i^* = y_i$ indicates the success of x_i. The human can observe the result produced by the AI when $d_i = \mathrm{AI}$, which is feedback for him/her to assess its reliability, but cannot when $d_i = \mathrm{human}$. p_i is the success probability of the AI for x_i. i is incremented when x_i is completed.

B. COLLABORATIVE CAPTCHA TASK

We developed a collaborative CAPTCHA (CC) task for training the reliance model and evaluating vPred-RC. Figure 2 shows a screenshot of the user interface. CAPTCHA is originally a task in which a human enters characters written in a noised and distorted image [19]. In the CC task, a worker can get assistance from a task AI that is trained to recognize characters in images.

Here, x_i is an RGB image of a CAPTCHA. y_i^* is a five-letter string of ground-truth labels for x_i. We prepared two verbal RCC candidates in addition to $\hat{c}^{\mathrm{w/o}}$: "I'm confident" and "I'm not confident," which are presented in a speech bubble from an image of the robot. Let us denote the two as $\hat{c}^{\mathrm{pos}}$ and $\hat{c}^{\mathrm{neg}}$, respectively.

$$\hat{\mathbf{c}} = \{\hat{c}^{\mathrm{pos}}, \hat{c}^{\mathrm{neg}}, \hat{c}^{\mathrm{w/o}}\}. \tag{1}$$

When $c_i = \hat{c}^{\mathrm{w/o}}$, the UI shows only the image of the robot and no speech bubble.

A worker first decides d_i (Fig. 2). If s/he chooses "AI," the task AI automatically enters its answer in a text box. The worker can observe the AI's answer before sending it to the host server but cannot edit it. If s/he chooses "Yourself," an empty text box appears, and s/he is asked to enter the characters. The worker repeats this 60 times.

The CC task refers to a sort of human-AI collaboration task such as human-robot collaborative picking [20], in which a human works with an intelligent robot that picks and places objects using a visual object-recognition method [21]. In this task, the human decides whether to perform the task by him/herself or ask the robot. Robot performance depends on the object recognition accuracy, which may change for various reasons. For example, it may not be able to recognize objects when they were not in its training dataset or when

Algorithm 1 Procedure for Selecting RCC

```
Require: $p_i$: AI success probability, $\hat{\mathbf{c}}$: set of RCC candidates, $\theta$: threshold to control the number of RCCs.
for $\hat{c} \in \hat{\mathbf{c}}$ do
    Predict $r^{\hat{c}_i}$ with reliance model.
    $\Delta^{\hat{c}_i} \leftarrow |r_i^{\hat{c}_i} - p_i|$.
end for
// Search best $\hat{c}$.
$\hat{c}_i^{\mathrm{best}} \leftarrow \underset{\hat{c} \in \hat{\mathbf{c}}, \hat{c} \neq \hat{c}^{\mathrm{w/o}}}{\operatorname{argmin}} \Delta_i^{\hat{c}}$
// Decide whether to provide $\hat{c}$.
if $\Delta^{\hat{c}^{\mathrm{w/o}}} - \Delta^{\hat{c}_i^{\mathrm{best}}} < \theta$ then
    return $\hat{c}^{\mathrm{w/o}}$
else
    return $\hat{c}_i^{\mathrm{best}}$
end if
```

environmental conditions change. To rely on the robot, the co-working human needs to understand what the robot can/cannot recognize in context. By selectively providing RCCs, the human can estimate the robot's ability with fewer communication costs.

C. VPRED-RC

vPred-RC adaptively selects whether to provide an RCC. Algorithm 1 describes the proceedure of the RCC selection and figure 1 graphically illustrates it. The main idea of vPred-RC is that it aims to avoid a discrepancy between the human reliance rate and the AI's success probability. For example, if an AI is likely to fail at a task, but a human is likely to rely on an AI without an RCC, it may be better to provide an RCC.

The reliance rate r_i is the probability that the human will assign x_i to the AI. A human is assumed to decide whether to rely on an AI depending on the current task and the collaboration history. We can consider r_i for each $\hat{c} \in \hat{\mathbf{c}}$:

$$r_i^{\Theta} = P(d_i = \mathrm{AI}|x_{:i}, c_{:i-1}, c_i = \hat{c}, y^*_{:i}, y_{:i-1}, d_{:i-1}). \quad (2)$$

Variables with the subscript $*_{:i}$ represent the vector of the sequence $(*_1, *_2, .., *_i)$. The discrepancy $\Delta_i^{\hat{c}}$ is the difference between r_i and p_i when $\hat{c}$ is provided:

$$\Delta_i^{\hat{c}} = |r_i^{\hat{c}} - p_i|. \quad (3)$$

First, vPred-RC refers to $\Delta_i^{\hat{c}}$ to evaluate $\hat{c}_i (\neq \hat{c}^{\mathrm{w/o}})$; the $\hat{c}_i$ whose $\Delta_i^{\hat{c}}$ is the lowest is the best RCC. Let us denote this $\hat{c}_i^{\mathrm{best}}$:

$$\hat{c}_i^{\mathrm{best}} = \underset{\hat{c} \in \hat{\mathbf{c}}, \hat{c} \neq \hat{c}^{\mathrm{w/o}}}{\operatorname{argmin}} \Delta_i^{\hat{c}}. \quad (4)$$

Next, vPred-RC decides whether to provide $\hat{c}_i^{\mathrm{best}}$ or omit it. Equation 5 is the criterion for the decision:

$$c_i = \begin{cases} \hat{c}_i^{\mathrm{best}} & (\Delta_i^{\hat{c}^{\mathrm{w/o}}} - \Delta_i^{\hat{c}_i^{\mathrm{best}}} < \theta) \\ \hat{c}^{\mathrm{w/o}} & (\text{elsewise}). \end{cases} \quad (5)$$

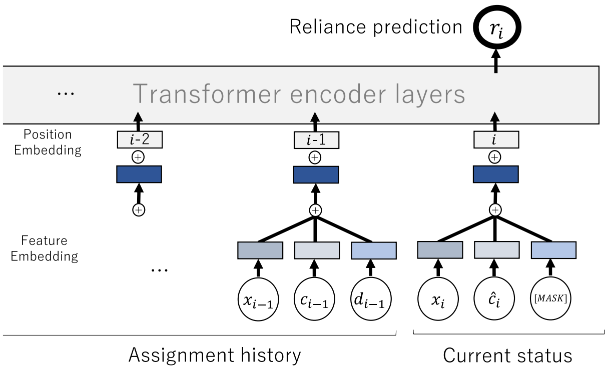

FIGURE 3. Reliance model.

θ represents the allowable range of $\Delta_i^{\hat{c}^{\mathrm{w/o}}}$ compared with $\Delta_i^{\hat{c}_i^{\mathrm{best}}}$ and controls how much vPred-RC omits RCCs. $\theta = 0$ means that vPred-RC omits c_i only when no RCC is predicted to be better rather than providing $\hat{c}_i^{\mathrm{best}}$, and increasing θ results in more omitted RCCs.

D. RELIANCE MODEL

The reliance model models the human cognition in deciding the assignment of a task and predicts $r_i^{\hat{c}_i}$. Figure 3 illustrates the structure of the model. It is based on the Transformer encoder [22], a deep-learning model that has shown great performance originally in natural language processing and increasingly been applied to other domains [23], [24]. By taking into account the collaboration history between a human and the target AI system, the reliance model can effectively capture human beliefs regarding an AI's capability.

The reliance model receives a history of collaboration between a human and AI $(x_{:i-1}, c_{:i-1}, d_{:i-1})$ and the current state (x_i, c_i). The history includes information such as when and to which task an RCC was provided and which decision the human made regarding the task, so the reliance model can capture a human's beliefs regarding what task they think the AI can execute to predict human decisions better.

In the implementation, x is a three-channel RGB array of a CAPTCHA image. $c \in \hat{\mathbf{c}}$ is a categorical label and represented as a one-hot vector. d is also a one-hot vector, an index of which is reserved for the case where it is hidden. Specifically, we hide d_i because they are not obtained when predicting r_i.

xs in the collaboration history and the status of the current task are first transformed into a one-dimensional vector with convolutional neural network layers. Other features are embedded with perceptrons. The embedded vectors are summed up with position embeddings, which give index information [22]. Then, the vectors are transformed by the Transformer encoder model, and a multi-layer perceptron predicts r_i from the transformed vector of the index i. Unlike equation 2, the reliance model cannot access y^* because we assume that the AI is not perfect. The detailed implementation

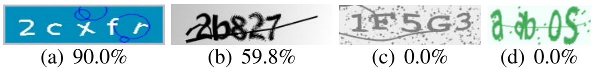

FIGURE 4. Examples from CAPTCHA datasets. We used left two datasets for training of task AI. Sub-caption shows accuracy of task AI for each dataset.

of the inputs of the reliance model should be modified depending on the nature of the target task. For example, if an AI can judge whether its task execution is successful or get feedback on its task execution, it may be better to include the information to predict human reliance more accurately.

The reliance model is trained in a supervised manner. We adopted a binary cross-entropy loss function for the training:

$$L = -\delta(d_i, \mathrm{AI}) \cdot \log(r_i) - \delta(d_i, \mathrm{human}) \cdot \log(1 - r_i), \quad (6)$$

where $\delta(a, b) = 1$ when $a = b$ and 0 when $a \neq b$.

IV. TRAINING RELIANCE MODEL

A. TASK IMPLEMENTATION

1) CAPTCHA DATASET AND TASK AI

Figure 4 shows examples of CAPTCHA images used in our experiments. We acquired four datasets[1] from Kaggle, a web platform for data scientists and machine learning practitioners. We split each dataset for training and testing. We excluded two datasets for training the task AI to replicate a bias in AI capability. For workers, understanding bias can help improve task assignment and result in fewer RCC requirements. Figure 4 also shows the accuracy of the task AI. The accuracy is actually biased by the dataset used for the training.

Each CAPTCHA image has five characters, and the task AI outputs the probability distribution that the j-th character $x_{i,j}$ is $\iota \in I$, where I is a set of alphabetic and numerical characters.

$$\mathrm{TaskAI}(x_{i,j}, \iota) = P(x_{i,j} = \iota). \quad (7)$$

When $d_i = \mathrm{AI}$, y_i is a sequence of the most probable $\iota \in I$ for each $x_{i,j}$.

$$y_i^{\mathrm{AI}} = \{\mathrm{argmax}_\iota(\mathrm{TaskAI}(x_{i,j}, \iota))\}_{j=1}^5. \quad (8)$$

The task AI was implemented using ResNet-18, a deep-learning model commonly used for image processing.

2) AI SUCCESS PROBABILITY

The confidence rate was calculated on the basis of the probability distribution output from the task AI [25].

$$conf \propto \Pi_{j=1}^5(\max_{\iota \in I}(\mathrm{TaskAI}(x_{i,j}, \iota))). \quad (9)$$

conf becomes higher the more probability there is that the task AI assigns to the most probable character. p_i was calculated on the basis of *conf* using a logistic regression model, which was trained to predict whether y_i^{AI} matches y_i^* from training datasets.

B. RELIANCE DATASET ACQUISITION

We made a *reliance dataset* to train the reliance model and evaluate vPred-RC. 228 participants were recruited with compensation of 100 JPY through Yahoo! Japan crowdsourcing. The data acquisition was conducted on a website. The participants were first provided pertinent information, and all participants consented to the participation. We instructed them on the flow of the CC task and asked five questions to check their comprehension of the task. 52 participants, who failed to answer the questions correctly, were excluded from the task. 145 participants completed the task (46 female, 98 male, one did not answer; aged $19 - 81$, $M = 48.4$, $SD = 13.1$). The protocol of the reliance dataset acquisition and the evaluation of vPred-RC was approved by the ethics committee of National Institute of Informatics.

x_i was randomly chosen for each participant from the test sub-datasets. We manipulated how many images to use from each CAPTCHA dataset so that the task AI's overall accuracy became 50% while keeping the task AI's accuracy for each dataset the same to avoid extreme over/under-reliance.

Whether to provide an RCC was randomly decided for each participant. The percentage of times that RCCs were provided was controlled to be 0, 20, 40, 60, 80, or 100%. We prepared two strategies to choose which RCC ($\hat{c}^{\mathrm{pos}}$ and $\hat{c}^{\mathrm{neg}}$) to show because we did not have the reliance model yet and could not use equation 4. For the *threshold* strategy, we chose $\hat{c}^{\mathrm{pos}}$ when p_i was higher than a threshold, which was determined by referring to the point closest to the top-left corner of the ROC[2] curve [26]. We set the threshold to 0.4475. For the other strategy, *proportional*, we chose a cue with a selection probability proportional to p_i. That is, $\hat{c}^{\mathrm{pos}}$ has a higher probability of being chosen if p_i is higher, but there is still a possibility that $\hat{c}^{\mathrm{neg}}$ could be chosen probabilistically as well, though with a lower probability than $\hat{c}^{\mathrm{pos}}$. We prepared the proportional strategy to broaden the variety in the dataset distribution by introducing randomness. 95 and 50 participants were assigned to the threshold and proportional conditions, respectively.

C. ANALYSIS OF RELIANCE DATASET

We briefly analyzed the reliance dataset to determine the trends for human-AI collaboration in the CC task. Figure 5 shows the distribution of the F-score for each participant. Here, the F-score was calculated with the number of times a participant assigned a task to the AI when the AI succeeded and assigned themselves when the AI failed. It reflects how much participants could predict the AI's performance with selective RCCs. The X-axis (rate) expresses the number of times that RCCs were provided.

[1] https://www.kaggle.com/datasets/utkarshdoshi/captcha-dataset
https://www.kaggle.com/datasets/alizahidraja/captcha-data
https://www.kaggle.com/datasets/kaushikmetha/captcha-images
https://www.kaggle.com/datasets/greysky/captcha-dataset

[2] Receiver Operating Characteristic.

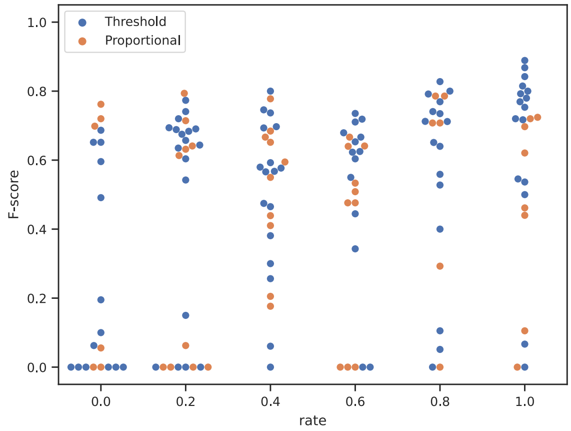

FIGURE 5. F-scores in reliance dataset.

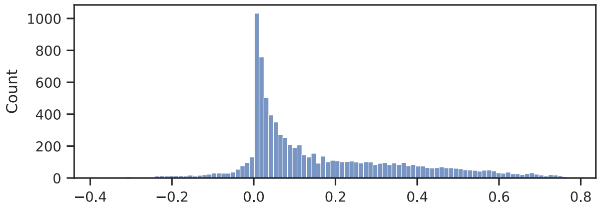

FIGURE 6. Distribution of $\Delta_i^{\hat{c}_i^{\text{best}}} - \Delta_i^{\hat{c}^{\text{w/o}}}$ for reliance dataset and thresholds for target percentage.

In general, the F-score decreased as the number of RCCs decreased. Notably, participants with a score of zero were concentrated in cases where the rate was 0.0 or 0.2. This is mainly because of under-reliance; the participants performed the task mostly by themselves and seldom used the AI.

The scores with the threshold strategy were slightly higher than proportional, but an ANCOVA (Analysis of covariance) with the rate variable as covariant showed no significant difference in the strategies ($F(1, 141) = 1.647; p = .20; \eta_p^2 = .012$). A summary table of the ANCOVA is available in Appendix (table 2).

D. TRAINING RESULTS

With the reliance dataset, we trained the reliance model and investigated its accuracy. The hyperparameters for the training are available in Appendix (table 4). Three-fold cross validation was performed with stratification of the data so that the percentage of the number of provided RCCs was aligned. We trained the model for 30 epochs. As a result, the maximum accuracy was 83.5% (SD=1.1%) on average at the 18th epoch. Hereafter, we combined the outputs of the three models through averaging as the reliance model to get the benefit of ensemble learning.

We further investigated $\Delta_i^{\hat{c}_i^{\text{best}}} - \Delta_i^{\hat{c}^{\text{w/o}}}$ that the model calculated with the dataset. This difference is used by vPred-RC to judge whether to provide an RCC (Equation 5). Figure 6 shows the distribution of the difference. Generally, the difference was positive, which means that it was predicted that it would be better to present $\hat{c}^{\text{best}}$ than omitting it to reduce Δ. However, in 8.0 percent of cases, the reliance model predicted that omitting an RCC instead would contribute to reducing the discrepancy.

V. EVALUATION

A. AIM

We evaluated whether vPred-RC can selectively provide RCCs at an effective timing. More specifically, we investigated whether vPred-RC can let users rely on an AI when it succeeds a task and avoid relying when fails.

B. PROCEDURE

The CC task was used to evaluate vPred-RC. The participants performed the task in a similar way as the reliance dataset acquisition. The difference is that it was vPred-RC that determined whether to provide RCCs by referring to each participant's decision-making history, whereas this was randomly determined in the reliance dataset acquisition. vPred-RC predicted the user reliance rate with the reliance model. 170 crowdworkers, none of whom participated in the data acquisition for the reliance dataset, were recruited for this experiment with compensation of 100 JPY. Using the comprehension checking questions, 55 participants were excluded from the CC task. 98 participants completed the task (38 female, 55 male, 5 did not answer; aged 17-86, M=47.4, SD=13.7). After the CC task, we asked the participants to answer Likert-scale questions to supplementarily investigate their subjective perception. We also asked them to freely comment on their experience with the task.

We compared the F-score for the humans' decisions in the vPred-RC condition with the random condition from the results of the threshold strategy in the reliance dataset. From the results of the reliance dataset acquisition and our preliminary experiments, we considered under-reliance to occur regardless of the choice of RCCs when the rate of the number of RCCs was below 20%. Therefore, we set θ so that the rate would be 30, 40, 50, 60, or 70% (See table 5 in Appendix), and let us denote the actual rate as a. Since θ cannot precisely control the rate, we used the results only in cases of $0.2 < a \leq 0.8$, and 18 and 13 participants were excluded because $a \leq 0.2$ and $0.8 < a$, respectively. Ignoring the results of $a = 0.2$ was disadvantageous for vPred-RC because the zero-score results of the random condition were not taken into account. Finally, 67 participants remained.

C. HYPOTHESIS

We hypothesized that by properly selecting whether to provide RCCs and which RCC to provide, vPred-RC earns higher F-score than the random condition when compared at the same rate.

D. RESULTS

Figure 7 illustrates the F-score for the humans' decisions. We conducted an ANCOVA to statistically analyze the

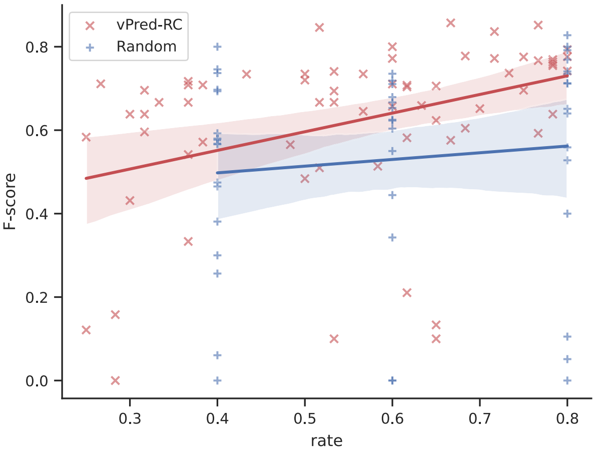

FIGURE 7. F-score of humans' decisions. Error bands shows 95% CI for linear regressions.

i	1	2	3	4	5	6
CAPTCHA						
AI correct?	false	true	false	false	true	false
RCC	w/o	pos	w/o	w/o	pos	w/o
Assignment	SELF	AI	SELF	SELF	AI	AI
Assignment correct?	✓	✓	✓	✓	✓	✗

(a)

i	1	2	3	4	5	6
CAPTCHA						
AI correct?	true	true	true	false	true	true
RCC	w/o	w/o	neg	pos	w/o	pos
Assignment	AI	SELF	SELF	AI	SELF	SELF
Assignment correct?	✓	✗	✗	✗	✗	✗

(b)

FIGURE 8. Successful and unsuccessful examples of vPred-RC.

results. There were significant effects for the number of cues ($F(1, 110) = 7.629; p = .0067; \eta_p^2 = .065$), the RCC selection method ($F(1, 110) = 6.575; p = .012; \eta_p^2 = .056$). The interaction effect was not statistically significant ($F(1, 110) = 1.387; p = .242; \eta_p^2 = .012$). This suggests that vPred-RC gets F-score per the number of RCCs more than the random condition in this range and supports the hypothesis. Therefore, we conclude that vPred-RC can reduce the number of times RCCs are provided while avoiding over/under-reliance by evaluating the effect of each RCC on the basis of human reliance prediction.

VI. DISCUSSION

A. EXAMPLES OF VPRED-RC'S BEHAVIOR

Figure 8(a) shows a successful example of vPred-RC. We need to mention that it is difficult to follow the actual dynamics of the interaction among vPred-RC, participants, and tasks, so our explanations here are post-hoc.

x_1, x_3, x_6 were from the datasets that were not used for the training of the task AI, and the AI could not correctly recognize them. vPred-RC first decided not to present a cue that the AI could not recognize, and the participant properly avoided assigning it to the AI. This was a risky but probabilistically possible decision. Only 35.6% of the participants assigned a task to the AI when $c_1 = \hat{c}^{\mathrm{w/o}}$, whereas 13.7% assigned a task to the AI when $c_1 = \hat{c}^{\mathrm{neg}}$. Providing $\hat{c}^{\mathrm{neg}}$ was judged less effective with equation 5. x_2 was from the training dataset, and the AI could recognize it correctly. vPred-RC provided $\hat{c}^{\mathrm{pos}}$ and successfully let the participant assign it to the AI. vPred-RC decided not to provide RCCs for $i = 3, 4$, and the participant properly avoided assigning them to the AI. Notably, although $p_4 = .489$ was above the threshold of the ROC curve (subsection IV-B), vPred-RC cancelled presenting $\hat{c}^{\mathrm{pos}}$. This example may suggest the potential of vPred-RC in borderline cases where whether the AI succeeds or fails is uncertain. vPred-RC provided $\hat{c}^{\mathrm{pos}}$ for x_5 and the participant could assign it to the AI. When $i = 6$, no RCC was provided and the participant falsely assigned it to the AI. A possible interpretation for not presenting the RCC is that the participant did x_1 and x_3, which were from the same dataset as x_6, and was expected to do it her/himself again. However, it is also possible to counter-argue that the participant had never observed whether the AI was capable of recognizing them, so it was likely that the participant tried to assign it to the AI.

Figure 8(b) illustrates an unsuccessful example. The target RCC rate for the participant was 30%, so vPred-RC needed to calibrate her/his reliance with a small number of RCCs. We should say that the success of the first trial was luck-based because, as we mentioned in the succesful case, the reliance is low when $c_1 = \hat{c}^{\mathrm{w/o}}$. Here, p was high (.700), but vPred-RC could not choose $\hat{c}^{\mathrm{pos}}$ because of a large θ. The participant did not assign x_2 to the AI, though s/he observed the successful result for x_1. This assignment does not follow a probabilistic expectation. 87.5% of the participants who observed AI success without an RCC when $i = 1$ assigned x_2 to the AI again. The failures for x_3 and x_4 were due to poor predictions of the AI success probability. p_3 and p_4 were.431 and.544, respectively, where the AI succeeded in the former and failed in the latter. Particularly, the failure for x_4 may have led the participant to under-trust, and s/he did not rely on the AI in spite of the presentation of $\hat{c}^{\mathrm{pos}}$.

B. SUBJECTIVE MEASURES AND COMMENTS

Although the main focus of this paper is reliance, the objective aspect of the collaboration, we supplementarily asked nine Likert-scale questions, with an expectation that appropriately selected RCCs would have a positive effect on the participants' subjective measures. Table 1 shows the questions, which were grouped into four categories. The "understanding of AI" category aimed to ask about the participants' subjective understanding of the performance of the AI gained through the trials. The "evaluation of selective RCCs" questions aimed to ask how useful the participants perceived the selectively provided RCCs to be. The "overall measures of system" questions aimed to ask the participants for their overall evaluation of the AI through the interaction in the CC task. Additionally, we asked the "understanding of

TABLE 1. Questions for subjective measures.

Category	Label	Question
Understanding of task	Instruction 1	Could you understand the instruction given before the task?
	Instruction 2	Could you understand the instruction after completing the task?
Understanding of AI	Performance understanding	How much do you understand the performance of the robot now?
	Performance predictability	How much can you predict whether the robot would succeed or fail now?
Evaluation of selective RCCs	RCC useful for assignment	Were the robot's messages useful to decide whether to assign a task to the robot?
	RCC useful for understanding performance	Were the robot's messages useful to understand the performance of the robot?
Overall measures of system	AI useful	Was the robot's support useful for this task?
	AI trustable	Do you think that the robot is trustable as a support robot?
	Want to use again	Are you willing to use the robot again if you join this task next?

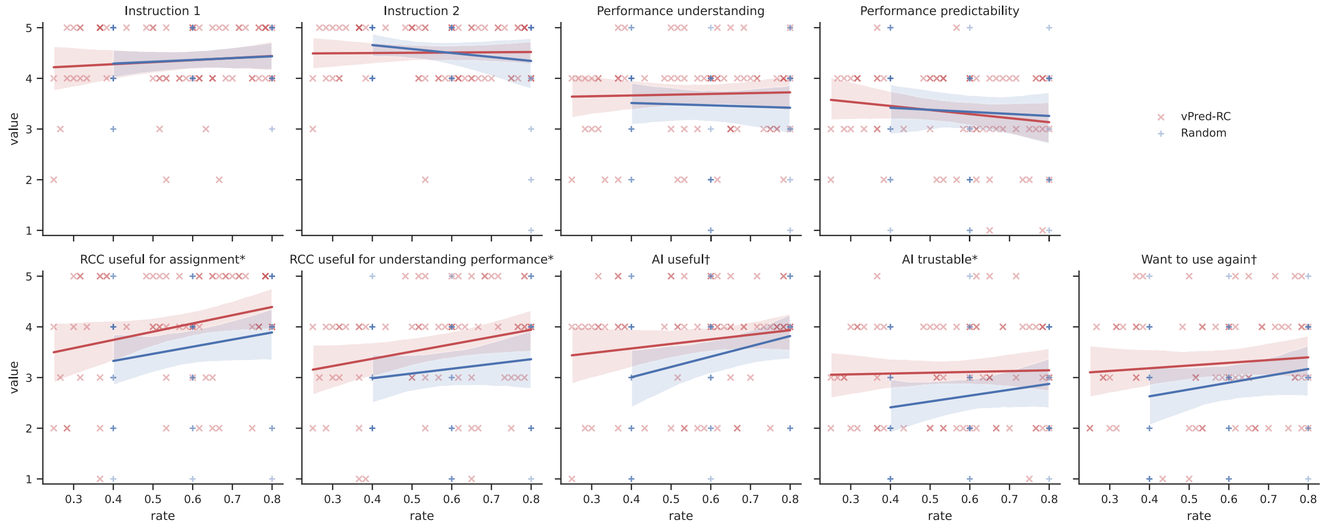

FIGURE 9. Results for subjective measures. * in subcaption indicates statistical significance between vPred-RC and Random ($p < .05$), and † indicates marginal significance ($p < .10$).

task" questions to check whether the human-AI collaboration setting in this paper was understandable for the participants.

We also prepared an open question that asked the participants to comment about what they thought during the task and how they decided task assignments. We acquired valid comments from 60 of the participants.

1) SUBJECTIVE MEASURES

Figure 9 shows the results of the Likert-scale questions. We conducted two-way ANCOVAs with the RCC selection methods as the independent variable and the rate as a covariant to analyze the difference between vPred-RC and the random condition. Summary tables of the ANCOVAs are available in Appendix (table 2). Here, we summarize the results for each category.

a: UNDERSTANDING OF TASK

Most of the participants evaluated their understandings of the task highly. There was no statistically significant effect of any variables for both questions.

b: UNDERSTANDING OF AI

There was no significant effect of the condition, rate, and their interaction effect for the two questions. Notably, the rate variable did not have an effect on the subjective evaluations. A possible reason is that the length of the task (60 trials) was enough to reach the upper bound of understanding the AI. There is a limit to the amount of information that RCCs can provide for several reasons, such as the prediction of the AI's success not being perfect and the performance of the AI not being stable even when a CAPTCHA is taken from a known dataset.

c: EVALUATION OF SELECTIVE RCCS

Significant effects of the condition were found for the two questions. That is, the participants perceived RCCs selected by vPred-RC to be more useful than those selected randomly when rates were aligned. The evaluation of RCCs decreased as the rate decreased, which is a natural result because the participants had less opportunity to get the benefit of RCCs.

d: OVERALL MEASURES OF SYSTEM

A significant effect of the condition was found for the "AI trustable" question, and the effects were marginally significant for the other questions. The scores of vPred-RC were consistent with the value of the rate for the last two questions, whereas the random condition got lower scores as the rate decreased. This suggests that, in the random condition, an RCC may not have been presented at the time the user needed it or that the RCC presented was inappropriate for participants, while vPred-RC could prevent the decrease in the scores.

In conclusion, RCCs selected by vPred-RC were more positively evaluated than the random condition. The participants thought that RCCs from vPred-RC were more useful. In addition, vPred-RC could prevent a decrease in trust even with a reduced number of RCCs.

2) COMMENTS

21 participants mentioned that they had focused on specific visual features such as "blue background color" or "crossed line" to decide to whom to assign tasks, suggesting that they were aware of the bias of the task AI's success probability, and most of them successfully captured the characteristics of the CAPTCHA datasets. 25 participants mentioned the provided RCCs. 14 participants explained that they assigned a task to the AI when $\hat{c}^{\text{pos}}$ was presented, and 8 explained that they avoid using the AI when $\hat{c}^{\text{neg}}$ was presented. One commented that he could not trust the AI when an RCC was missing. One provided a negative comment on the provided RCCs because the AI failed even though they provided $\hat{c}^{\text{pos}}$. This may indicate distrust in RCCs. The problem of distrusted RCCs may need to be handled in ways such as apologies, excuses, or explanations and dialogues, which are found to be effective for trust repair [27], [28], [29], [30], [31].

C. SCOPE AND LIMITATIONS

Our formalization in subsection III-A assumes a situation in which a human successively decides whether to assign a task to an AI system. In other words, the reliance model aims to capture the dynamics of human-AI collaboration *between* trials. This assumption is likely to suit tasks with characteristics such as:

- The human can allow for trial and error, so it is less necessary for him/her to continuously monitor an AI's behavior in order to intervene (for example, picking a solid object, cleaning floors).
- Each trial in a task is performed in a short period of time (X-ray inspection).
- Total performance is more important than the accuracy for each trial (recommending items to customers).

However, vPred-RC is not immediately applicable to tasks in which we need to consider the dynamics *during* each trial. For example, when it comes to autonomous cars, an AI system's action may immediately lead to a critical failure. Particularly, level 2 or 3 autonomous cars require users to constantly monitor the situation and system behavior to intervene when necessary, which makes the dynamics during a trial more important. We are currently tackling the problem of the during-task setting by using this paper's approach of selecting RCCs by predicting their effect on human reliance. Here, we are considering an autonomous driving monitoring task in which a user judges whether to rely on an autonomous car or drive themselves while monitoring the car's object detection results as RCCs [32]. However, the dynamics between trials are still important for the long-term use of the system, and combining the two approaches has the potential to further deepen the relationship between humans and AIs. Another limitation is that vPred-RC assumes that a task's result can be categorized as either success or failure. We also need further consideration for tasks whose results should be evaluated continuously.

The participants recruited from the crowdsourcing platform had a diverse set of demographic characteristics, and from this perspective, we believe our data to have good representativeness. We also conducted a post-hoc power analysis on the ANCOVA in subsection V-D using G*Power [33] and calculated a power $(1 - \beta)$ of 0.73. However, there are also potential limitations such as the nationality of the participants, as most participants on the platform are Japanese. A potential bias in the experimental setup is the difficulty of the task for both a user and an AI. That is, which CAPTCHA image a person/AI can read which they cannot. This paper assumes a situation in which a user has a better capability of reading the images but uses the AI for efficiency. However, if the legibility of an image for the user is too low, s/he may need to rely on an AI even when the AI performance is also low, which can make the optimal strategy of RCC selection much more complex.

Our experimental results that demonstrate the success of vPred-RC indicate that 83.5% accuracy is fair enough for vPred-RC to address the problem of reliance calibration with fewer RCCs on average. However, as we demonstrated in subsection VI-A, there were also failures caused by poor prediction of the reliance model, which means room for further improvement.

We aimed to change the value of θ to manipulate the rate of the number of RCCs, and we investigated whether vPred-RC could actually manipulate it as intended. The correlation coefficient between the target and actual rates at which RCCs were provided was .679 ($p < .0001$), which suggests that we can control the number of RCCs to some extent by changing θ. In actual use, however, we need to consider the trade-off between collaboration performance and the communication cost of RCCs rather than rigidly target the number of RCCs. A future direction for this work is to integrate machine-learning methods to adjust θ. A possible approach is using reinforcement learning (RL), in which another model whose structure is the same as the reliance model but it learns not r_i but *theta* with a reward function that balances the performance and cost.

Our formalization does not consider human capability for a task. Two participants commented that they used the task AI when they were not confident in their answers, and one said that the AI was useless because it could not recognize CAPTCHA that she cannot read. In the CC task, humans were not perfect as well (81.7% accuracy when d_i = human). While our experiments successfully demonstrated that vPred-RC can effectively calibrate human reliance with a measure of how many times humans assign a task to the AI if and only if the AI can succeed, to improve the total collaboration performance, we still need to take into account the capability of a human and compare it with that of an AI.

VII. CONCLUSION

To address the problem of excessive amounts of information from XAI, this paper proposed vPred-RC. It dynamically selects which RCC to provide by predicting its effect on human reliance using an AI reliance model that predicts the probability of a human assigning a task to an AI for cases in which an RCC is provided or not. vPred-RC aims to avoid a discrepancy between the task success probability of an AI and the human reliance rate.

vPred-RC was extended from our previously proposed algorithm, Pred-RC. The main difference in the extension is that it handles verbal cues as RCCs toward interactive human-robot collaboration, whereas Pred-RC considers a situation in which an AI system displays an indicator of its confidence rates as RCCs.

We tested vPred-RC for a human-AI collaboration task. First, the reliance model was found to predict human decision-making with 83.5% accuracy. Next, we found that the RCCs selected by vPred-RC enabled participants to more accurately assign tasks to an AI when and only when the AI succeeded compared with randomly selected ones, suggesting that vPred-RC can successfully calibrate human reliance with a reduced number of RCCs. Last, we revealed that vPred-RC's selective RCCs acquire better subjective evaluations than random ones. These results demonstrate the potential of vPred-RC for human-robot collaboration such as human-robot collaborative picking (subsection III-B). With vPred-RC, humans are expected to efficiently understand what a robot can and cannot do, leading to making full use of it with fewer communication costs.

APPENDIX. DATA AND CODE AVAILABILITY

The implementation of vPred-RC, the reliance dataset, and the results of the evaluation are available at https://github.com/fuku5/vPred-RC.

APPENDIX. ANCOVA RESULTS

A. RELIANCE DATASET

Table 2 shows a summary table of the ANCOVA for the reliance dataset analysis (subsection IV-C).

TABLE 2. Reliance dataset analysis.

	df	F	p	η_p
condition	1	1.6467	0.2015	0.0115
rate	1	12.8807	0.0005***	0.0837
rate:condition	1	1.3755	0.2428	0.0097

TABLE 3. Evaluation of vPred-RC.

(a) Instruction 1

	df	F	p	η_p^2
condition	1	0.0009	0.9761	0.0000
rate	1	0.9486	0.3322	0.0086
rate:condition	1	0.0031	0.9559	0.0000

(b) Instruction 2

	df	F	p	η_p^2
condition	1	0.0005	0.9827	0.0000
rate	1	0.5008	0.4807	0.0045
rate:condition	1	1.0116	0.3167	0.0091

(c) Performance understanding

	df	F	p	η_p^2
condition	1	1.4315	0.2341	0.0128
rate	1	0.0000	0.9944	0.0000
rate:condition	1	0.1279	0.7213	0.0012

(d) Performance predictability

	df	F	p	η_p^2
condition	1	0.0346	0.8528	0.0003
rate	1	1.2988	0.2569	0.0117
rate:condition	1	0.1215	0.7281	0.0011

(e) RCC useful for assignment

	df	F	p	η_p^2
condition	1	3.9477	0.0494*	0.0346
rate	1	5.5109	0.0207*	0.0477
rate:condition	1	0.0956†	0.7577	0.0009

(f) RCC useful for understanding AI performance

	df	F	p	η_p^2
condition	1	4.2278	0.0421*	0.0370
rate	1	3.3975	0.0680†	0.0300
rate:condition	1	0.3265	0.5689	0.0030

(g) AI useful

	df	F	p	η_p^2
condition	1	3.0444	0.0838†	0.0269
rate	1	5.0823	0.0262*	0.0442
rate:condition	1	0.8474	0.3593	0.0076

(h) AI trustable

	df	F	p	η_p^2
condition	1	6.0794	0.0152*	0.0524
rate	1	0.9756	0.3255	0.0088
rate:condition	1	0.7297	0.3949	0.0066

(i) Want to use again

	df	F	p	η_p^2
condition	1	3.7007	0.0570†	0.0325
rate	1	1.9920	0.1610	0.0178
rate:condition	1	0.4197	0.5185	0.0038

B. EVALUATION

Tables 3 show summary tables of the ANCOVAs for the subjective measures (subsection VI-B).

TABLE 4. Hyperparameters for training reliance model.

hyperparameter	value
number of Transformer-encoder layers	3
number of Transformer-encoder heasds	16
dimension of Transformer-encoder input	128
dimension of Transformer-encoder feedforward network model	2048
dropout rate	0.2
hidden sizes for multi-layer perceptron	[1024, 1024, 1024]

TABLE 5. θ for evaluation experiment.

target percentage	θ
30%	0.25341
40%	0.15557
50%	0.09778
60%	0.05829
70%	0.03341

APPENDIX. ENVIRONMENT AND PARAMETERS

For the implementation, we used following softwares: python 3.8.10, torch 1.13.0, and numpy 1.23.2. Table 4 shows the hyperparameters of the reliance model. Table 5 shows the values of θ used for the evaluation.

REFERENCES

[1] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[2] A. Rai, "Explainable AI: From black box to glass box," *J. Acad. Marketing Sci.*, vol. 48, no. 1, pp. 137–141, Jan. 2020.

[3] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 57, no. 3, pp. 407–434, May 2015.

[4] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 39, no. 2, pp. 230–253, Jun. 1997, doi: 10.1518/001872097778543886.

[5] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *Proc. 11th ACM/IEEE Int. Conf. Human Robot Interact. (HRI)*, Mar. 2016, pp. 101–108.

[6] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman, "A design methodology for trust cue calibration in cognitive agents," in *Virtual, Augmented and Mixed Reality: Designing and Developing Virtual and Augmented Environments*, R. Shumaker and S. Lackey, Eds. Cham, Switzerland: Springer, 2014, pp. 251–262.

[7] A. N. Ferguson, M. Franklin, and D. Lagnado, "Explanations that backfire: Explainable artificial intelligence can cause information overload," in *Proc. Annu. Meeting Cognit. Sci. Soc.*, vol. 44, no. 44, 2022, p. 3817.

[8] C. Oh, J. Song, J. Choi, S. Kim, S. Lee, and B. Suh, "I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence," in *Proc. Conf. Human Factors Comput. Syst. (CHI)*. New York, NY, USA: Association for Computing Machinery, 2018, pp. 1–13.

[9] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: A technique for explainable AI and its effects on human perceptions," in *Proc. 24th Int. Conf. Intell. User Interfaces*, Mar. 2019, pp. 263–274.

[10] Y. Fukuchi and S. Yamada, "Selectively providing reliance calibration cues with reliance prediction," in *Proc. Annu. Meeting Cognit. Sci. Soc.*, vol. 45, 2023, pp. 1–8. [Online]. Available: https://escholarship.org/uc/item/8zp6g0mj

[11] N. Scharowski, S. A. C. Perrig, N. von Felten, and F. Brühlmann, "Trust and reliance in XAI—Distinguishing between attitudinal and behavioral measures," in *Proc. CHI Workshop Trust Reliance AI-Human Teams*, 2022, pp. 1–6.

[12] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-aware decision making for human–robot collaboration: Model learning and planning," *ACM Trans. Human Robot Interact.*, vol. 9, no. 2, pp. 1–23, Jan. 2020, doi: 10.1145/3359616.

[13] T. Helldin, G. Falkman, M. Riveiro, and S. Davidsson, "Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving," in *Proc. 5th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.* New York, NY, USA: Association for Computing Machinery, Oct. 2013, pp. 210–217, doi: 10.1145/2516540.2516554.

[14] R. Häuslschmid, M. von Bülow, B. Pfleging, and A. Butz, "Supporting trust in autonomous driving," in *Proc. 22nd Int. Conf. Intell. User Interfaces* New York, NY, USA: Association for Computing Machinery, Mar. 2017, pp. 319–329, doi: 10.1145/3025171.3025198.

[15] J. M. McGuirl and N. B. Sarter, "Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information," *Human Factors, J. Human Factors Ergonom. Soc.*, vol. 48, no. 4, pp. 656–665, Dec. 2006.

[16] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," *PLoS One*, vol. 15, no. 2, Feb. 2020, Art. no. e0229132, doi: 10.1371/journal.pone.0229132.

[17] K. Okamura and S. Yamada, "Empirical evaluations of framework for adaptive trust calibration in human-AI cooperation," *IEEE Access*, vol. 8, pp. 220335–220351, 2020.

[18] M. Körber, E. Baseler, and K. Bengler, "Introduction matters: Manipulating trust in automation and reliance in automated driving," *Appl. Ergonom.*, vol. 66, pp. 18–31, Jan. 2018.

[19] L. von Ahn, M. Blum, N. J. Hopper, and J. Langford, "CAPTCHA: Using hard AI problems for security," in *Advances in Cryptology—EUROCRYPT*, E. Biham, Ed. Berlin, Germany: Springer, 2003, pp. 294–311.

[20] M. Rieder and N. Bartneck, "Demonstrator for a collaborative human–robot picking system," in *Intelligent Human Systems Integration*, vol. 22. New York, NY, USA: AHFE Open Access, 2022.

[21] Q. M. Marwan, S. C. Chua, and L. C. Kwek, "Comprehensive review on reaching and grasping of objects in robotics," *Robotica*, vol. 39, no. 10, pp. 1849–1882, Oct. 2021.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.

[23] S. Matsumori, K. Shingyouchi, Y. Abe, Y. Fukuchi, K. Sugiura, and M. Imai, "Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1878–1887.

[24] Y. Fukuchi, M. Osawa, H. Yamakawa, and M. Imai, "Explaining intelligent agent's future motion on basis of vocabulary learning with human goal inference," *IEEE Access*, vol. 10, pp. 54336–54347, 2022.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1321–1330.

[26] Z. H. Hoo, J. Candlish, and D. Teare, "What is an ROC curve?" *Emergency Med. J.*, vol. 34, no. 6, pp. 357–359, 2017. [Online]. Available: https://emj.bmj.com/content/34/6/357

[27] G. M. Lucas, J. Boberg, D. Traum, R. Artstein, J. Gratch, A. Gainer, E. Johnson, A. Leuski, and M. Nakano, "Getting to know each other: The role of social dialogue in recovery from errors in social robots," in *Proc. 13th ACM/IEEE Int. Conf. Human Robot Interact. (HRI)*. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 344–351.

[28] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *Proc. 15th ACM/IEEE Int. Conf. Human Robot Interact. (HRI)*. New York, NY, USA: Association for Computing Machinery, Mar. 2020, pp. 33–42, doi: 10.1145/3319502.3374839.

[29] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, Eds. Cham, Switzerland: Springer, 2015, pp. 574–583.

[30] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, "'I don't believe you': Investigating the effects of robot trust violation and repair," in *Proc. 14th ACM/IEEE Int. Conf. Human Robot Interact.* Piscataway, NJ, USA: IEEE Press, Mar. 2019, pp. 57–65.

[31] S. S. Sebo, M. Traeger, M. Jung, and B. Scassellati, "The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human–robot teams," in *Proc. 13th ACM/IEEE Int. Conf. Human Robot Interact. (HRI)*. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 178–186, doi: 10.1145/3171221.3171275.

[32] Y. Fukuchi and S. Yamada, "Selective presentation of AI object detection results while maintaining human reliance," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2023.

[33] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," *Behav. Res. Methods*, vol. 41, no. 4, pp. 1149–1160, Nov. 2009.

YOSUKE FUKUCHI was born in Japan, in 1994. He received the M.E. and Ph.D. degrees in engineering from Keio University, Yokohama, Japan, in 2019 and 2023, respectively.

From 2019 to 2021, he was an Assistant Professor with Keio University. From 2021 to 2022, he was a Project Researcher with the Keio Leading-Edge Laboratory of Science and Technology (KLL). He is currently a Project Researcher with the National Institute of Informatics, Tokyo, Japan. His research interests include human-AI interaction and artificial intelligence. He is a member of the Japanese Society for Artificial Intelligence.

SEIJI YAMADA (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in artificial intelligence from Osaka University. He is currently a Professor with the National Institute of Informatics and The Graduate University for Advanced Studies (SOKENDAI). Previously, he was with the Tokyo Institute of Technology. His research interests include the design of intelligent interaction, including human-agent interaction, intelligent web interaction, and interactive machine learning.