Selectively Providing Reliance Calibration Cues With Reliance Prediction

Yosuke Fukuchi[1] **(fukuchi@nii.ac.jp)**

Seiji Yamada[1,2] **(seiji@nii.ac.jp)**

[1] National Institute of Informatics, 2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
[2] The Graduate University for Advanced Studies (SOKENDAI), Shonan Vellage, Hayama, Kanagawa, 240-0115, Japan

Abstract

For effective collaboration between humans and intelligent agents that employ machine learning for decision-making, humans must understand what agents can and cannot do to avoid over/under-reliance. A solution to this problem is adjusting human reliance through communication using reliance calibration cues (RCCs) to help humans assess agents' capabilities. Previous studies typically attempted to calibrate reliance by continuously presenting RCCs, and when an agent should provide RCCs remains an open question. To answer this, we propose Pred-RC, a method for selectively providing RCCs. Pred-RC uses a cognitive reliance model to predict whether a human will assign a task to an agent. By comparing the prediction results for both cases with and without an RCC, Pred-RC evaluates the influence of the RCC on human reliance. We tested Pred-RC in a human-AI collaboration task and found that it can successfully calibrate human reliance with a reduced number of RCCs.

Keywords: reliance calibration; reliance prediction; human-AI collaboration

Introduction

Machine learning (ML) is a powerful tool for robots and agents that collaborate with humans. There have been many trials of introducing ML, and such AI agents have shown great performance in various fields (Ren & Bao, 2020; Gul, Rahiman, & Alhady, 2019; Kalashnikov et al., 2018). However, as ML models get more complex, it becomes more difficult for end users to understand how to adequately use AI agents (Adadi & Berrada, 2018; Rai, 2020), a consequence of which is that users over-rely or under-rely on them (Hoff & Bashir, 2015; Parasuraman & Riley, 1997). Over-reliance, in which a human overestimates the capability of an AI agent, can cause misuse and task failure (Robinette, Li, Allen, Howard, & Wagner, 2016). It even leads to even serious accidents, particularly for embodied agents such as robots and autonomous vehicles. Under-reliance is also problematic because it results in disuse, increases human workload, and degrades the total collaboration performance.

Previous studies attempted to adjust human reliance by providing signals or information elements used by humans to assess an AI's capability (de Visser, Cohen, Freedy, & Parasuraman, 2014). In this paper, we call them reliance calibration cues (RCCs). For example, presenting an AI model's confidence rate is shown to be effective for an RCC (McGuirl & Sarter, 2006; Zhang, Liao, & Bellamy, 2020).

This work was supported in part by JST, CREST, Japan, under Grant JPMJCR21D4.

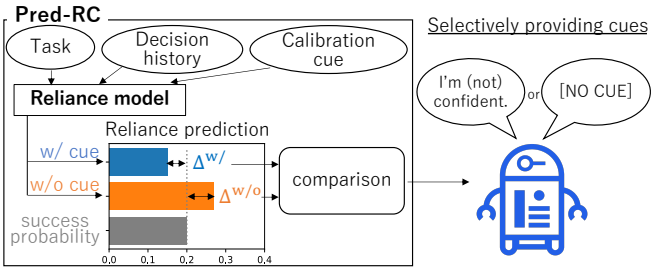


Figure 1: Predictive Reliance Calibrator (Pred-RC) enables AI system to selectively provide trust calibration cues. In Pred-RC, cognitive reliance model predicts probability that human will assign current task to AI. By considering both cases with and without cue provision, Pred-RC evaluates how much cue will contribute to trust calibration and decides whether to provide it.

A challenge facing reliance calibration with RCCs lies in the timing at which to provide them. In typical previous studies, RCCs are provided continuously, but this is not always realistic, for example when a robot verbally provides them. There is a trade-off between successful calibration and reducing the communication cost, but computational methods for deciding when to provide them are quite limited. Okamura & Yamada proposed a method for selectively providing RCCs by detecting over/under-reliance (Okamura & Yamada, 2020a, 2020b). However, their method simply checks a human's false assignment of past tasks to him/herself or an AI and does not capture a human's cognitive aspects such as his/her past experiences collaborating with AI and beliefs of what kind of tasks the AI can do.

This paper proposes *Predictive Reliance Calibrator (Pred-RC)*, a method for selectively providing RCCs (Fig. 1). The main idea of Pred-RC is that it selects 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. In Pred-RC, a cognitive reliance model predicts human reliance rates in both cases where an RCC is provided or not. By comparing the predicted rates with the success probability (actual reliability), Pred-RC evaluates the impact of an RCC for reliance calibration.

This paper reports an experiment on reliance calibration by

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Pred-RC using crowdsourcing. 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 did not decline with Pred-RC's selective RCCs, whereas that of workers whose RCCs were randomly provided got worse as fewer RCCs were provided, suggesting that Pred-RC enables an AI to selectively provide RCCs at the proper timing by predicting and comparing reliance with and without an RCC.

Background

Trust/reliance calibration

Reliance is a concept relevant to trust, and it is sometimes studied inclusively in the field of engineering. Trust is attitudinal and a psychological construct, while reliance focuses on the behaviors of humans, which is directly observable and thus an objective measure (Scharowski, Perrig, von Felten, & Brühlmann, 2022). 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, one of which is to change an agent's actions (Dubois & Ny, 2020; Sheng et al., 2021). For example, Chen et al. proposed trust-POMDP, a computational model that allows an AI to decide an action with awareness of human trust (Chen, Nikolaidis, Soh, Hsu, & Srinivasa, 2020). They demonstrated that with trust-POMDP, a robot automatically generates behavior that involves tackling an easier task first and successfully earns human trust.

Another approach is to explicitly provide information or communicative signals that help humans assess an AI's capability. Commonly used RCCs are the confidence rate or uncertainty of an AI's decision-making (McGuirl & Sarter, 2006; Zhang et al., 2020; Helldin, Falkman, Riveiro, & Davidsson, 2013). In this paper, Pred-RC provides confidence information as an RCC.

Some studies focused on continuously providing RCCs and demonstrated its effectiveness (Helldin et al., 2013; Häuslschmid, von Bülow, Pfleging, & Butz, 2017). McGuirl & Sarter compared providing dynamic system confidence with overall reliability only and found that the former can improve trust calibration (McGuirl & Sarter, 2006).

However, there are at least two potential concerns with the continuous provision of RCCs. One is that too many RCCs can be annoying for humans. This depends mainly on how the RCC are provided, and a typical negative case is that in which a robot verbally provides them. The other concern is that humans sometimes pay less attention to continuously displayed information. Okamura & Yamada found that, 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 (Okamura & Yamada, 2020a). Therefore, we aim to achieve successful reliance calibration by not continuously but selectively providing RCCs.

Very few studies have focused on a computational method for adaptively 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 (Okamura & Yamada, 2020a, 2020b). 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 Pred-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.

Reliance estimation

A basic idea of Pred-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.

Self-report trust scales are commonly used to measure trust (Jian, Bisantz, & Drury, 2000; Madsen & Gregor, 2000; Yagoda & Gillan, 2012). Some studies focus on neural metrics to infer human trust using fMRI or EEG (Fett, Gromann, Giampietro, Shergill, & Krabbendam, 2012; Choo & Nam, 2022). A weak point of these methods is that they can be intrusive during a task execution. A more relevant approach to this study focuses on human behavior. Walker et al. proposed a method for inferring human trust on the basis of their gaze movements (Walker, Verwey, & Martens, 2018). Human interventions or takeovers of a robot's action is an indicator of poor trust, and Muir incorporated them into human trust models (Krber, Baseler, & Bengler, 2018). Another factor is a human's decision-making regarding whether to assign a task to themselves or an AI (Okamura & Yamada, 2020a, 2020b), and we follow this approach.

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

Selectively providing trust calibration cues

Formalization

This paper formalizes human-AI collaboration with selectively provided RCCs as a tuple $(x, \hat{c}, c, d, y^*, y, p)$. 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. $\hat{c}_i$ is a potential RCC for the AI system when x_i is given, and Pred-RC decides whether to provide it. c_i is the RCC actually provided to the human:

$$c_i = \begin{cases} \hat{c}_i & \text{(if RCC is provided)} \\ [MASK] & \text{(elsewise)}. \end{cases} \quad (1)$$

$c_i = [MASK]$ means that no RCC is provided.

The human observes (x_i, c_i, y_i^*) and determines whether to assign x_i to him/herself or the AI agent. Let $d_i \in \{\text{AI}, \text{human}\}$ be the agent responsible for 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 $d_i = \text{AI}$, which is feedback for him/her to assess its reliability, but cannot when $d_i = \text{human}$.

p_i is the success probability of the AI for x_i. i is incremented when x_i is completed.

Pred-RC

Pred-RC adaptively selects whether to provide an RCC (Fig. 1). The main idea of Pred-RC is that it aims to avoid a discrepancy between the human reliance rate r_i and the AI's success probability p_i. Here, r_i is the probability that the human will assign x_i to the AI. Pred-RC considers two types of r_i:

$$\begin{aligned} r_i^{\text{w/}} &= P(d_i = \text{AI} | x_{:i}, c_{:i-1}, c_i = \hat{c}_i, d_{:i-1}, y^*_{:i}, y_{:i-1}), \\ r_i^{\text{w/o}} &= P(d_i = \text{AI} | x_{:i}, c_{:i-1}, c_i = [MASK], d_{:i-1}, y^*_{:i}, y_{:i-1}). \end{aligned} \quad (2)$$

The difference between $r^{\text{w/}}$ and $r^{\text{w/o}}$ is whether $\hat{c}_i$ is provided to the human or not. Variables with the subscript $*_{:i}$ represent the vector of the sequence $(*_1, *_2, .., *_i)$.

The discrepancy Δ_i is the difference between r_i and p_i:

$$\begin{cases} \Delta_i^{\text{w/}} &= |r_i^{w/} - p_i|, \\ \Delta_i^{\text{w/o}} &= |r_i^{w/o} - p_i|. \end{cases} \quad (3)$$

Pred-RC compares $\Delta_i^{\text{w/}}$ and $\Delta_i^{\text{w/o}}$ and decides whether to provide $\hat{c}_i$. Equation 1 is rewritten as follows:

$$c_i = \begin{cases} \hat{c}_i & (\Delta_i^{\text{w/o}} - \Delta_i^{\text{w/}} < threshold) \\ [MASK] & \text{(elsewise)}. \end{cases} \quad (4)$$

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

Reliance model

The reliance model is a cognitive model that predicts both $r_i^{\text{w/}}$ and $r_i^{\text{w/o}}$. Figure 2 illustrates the structure of the model. It is based on the Transformer encoder (Vaswani et al., 2017), a deep-learning model that has shown great performance originally in natural language processing and is being applied to various domains (Han et al., 2022; Mousavi, Ellsworth, Weiqiang, Chuang, & Beroza, 2020) including human-computer interaction (Matsumori et al., 2021; Fukuchi, Osawa, Yamakawa, & Imai, 2022). 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.

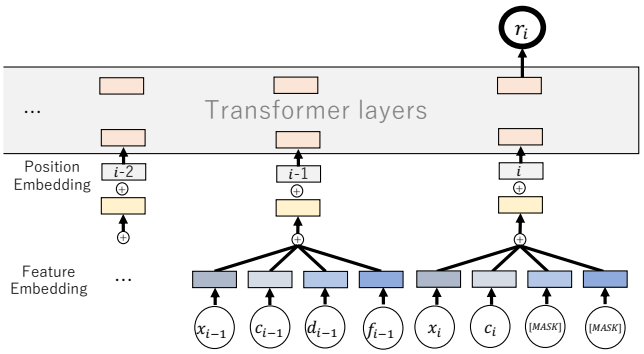


Figure 2: Structure of reliance model

The reliance model receives a history of collaboration between a human and AI $(x_{:i-1}, c_{:i-1}, d_{:i-1}, f_{:i-1})$ 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.

Each feature in the collaboration history is first embedded with perceptrons. The embedded vectors are summed up with position embeddings, which give index information (Vaswani et al., 2017). 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, although it is an important information for the human to assess the AI's reliability by comparing it with the AI's result. Instead of this, we included f, feedback from the human's task result:

$$f_i = \begin{cases} 0 & (d_i = \text{AI}) \\ 1 & (d_i = \text{human and } y_i \text{ matches the AI}'\text{s result}) \\ 2 & (d_i = \text{human and } y_i \text{ does not match the AI}'\text{s result}). \end{cases} \quad (5)$$

We masked d_i and f_i because they are not obtained when predicting r_i.

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

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

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

When inferring r_i, we run the reliance model with both cases in which $c_t = \hat{c}_t$ and $c_t = [MASK]$, the results of which are the predicted values of $r^{\text{w/}}$ and $r^{\text{w/o}}$, respectively.

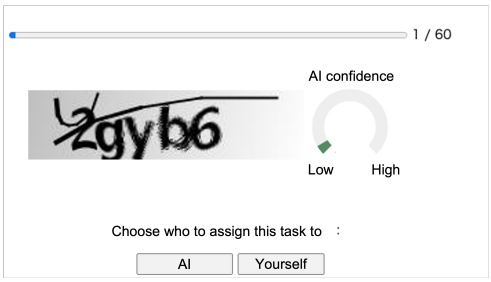
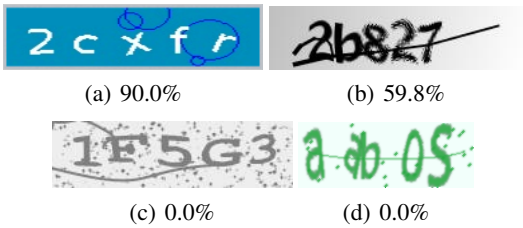


Figure 3: Screenshot of user interface for CC task



(a) 90.0% (b) 59.8%

(c) 0.0% (d) 0.0%

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

Training reliance model

Task

We developed a collaborative CAPTCHA (CC) task for training the reliance model and evaluating Pred-RC. Figure 3 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 (von Ahn, Blum, Hopper, & Langford, 2003). 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 image, and y_i^* is a ground-truth label for x_i. $\hat{c}_i$ is the confidence rate of the task AI for x_i.

A worker first chooses d_i (Fig. 3). 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.

Task implementation

CAPTCHA dataset and task AI Figure 4 shows examples of CAPTCHA images used in our experiments. We acquired four datasets from Kaggle, a web platform for data scientists and machine learning practitioners[1]. 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.

[1]https://www.kaggle.com/.

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.

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

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

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

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

RCC and success probability The confidence rate was calculated on the basis of the probability distribution output from the task AI (Guo, Pleiss, Sun, & Weinberger, 2017).

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

$\hat{c}_i$ 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 $\hat{c}_i$ using logistic regression. The logistic regression model was trained to predict whether y_i^{AI} matches y_i^* from training datasets.

We adopted the confidence rate, $\hat{c}_i$, rather than the success probability because, in our pilot experiment, we found that the confidence rate calibrates worker reliance better than the success probability. This is presumably because the success probability, which was calculated with logistic regression, was distributed steeply around 0% and 100%, whereas the distribution of the confidence rate was flatter.

Reliance dataset acquisition and model training

We made a *reliance dataset* to train the reliance model and evaluate Pred-RC. The dataset was composed of sequences of the tuple $(x, \hat{c}, c, d, y^*, y, p)$.

250 participants were recruited with compensation of 100 JPY through Yahoo! Japan crowdsourcing[2]. 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. 99 participants, who failed to answer the questions correctly, were excluded from the CC task, and 151 participants remained (60 female and 91 male; aged $20-78, M = 47.5, SD = 12.2$). The protocol of the reliance dataset acquisition and the evaluation of Pred-RC was approved by the ethics committee of the 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.

[2]https://crowdsourcing.yahoo.co.jp/

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 trained the reliance model and investigated its accuracy with the reliance dataset. We performed k-fold cross validation with stratification of the data to align the percentage of the number of provided RCCs. We set $k = 10$. After 50 epochs of training, the reliance model predicted d_i with a maximum accuracy of 81.6% (95% CI[3]: 80.0%, 83.2%) on average at the 25th epoch.

Evaluation

Aim

We evaluated whether Pred-RC can selectively provide RCCs at an effective timing. More specifically, we investigated whether Pred-RC can reduce the number of RCCs while avoiding over/under-reliance.

Procedure

The CC task was used to evaluate Pred-RC. The participants performed the task in a similar same way as the reliance dataset acquisition. The difference is that it was Pred-RC that determined whether to provide RCCs with each participant's decision-making history, whereas this was randomly determined in the reliance dataset acquisition. Pred-RC predicted the user reliance rate with the reliance model.

91 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, 39 participants were excluded from the CC task, and 52 participants remained (14 female and 38 male; aged 21-65, M=43.3, SD=10.2). After the CC task, we also asked the participants to freely comment about their experience with the task.

We prepared a *threshold* set to control the number of RCCs to provide. This was determined by referring to the distribution of $\Delta_i^{\mathrm{w/o}} - \Delta_i^{\mathrm{w/}}$ calculated with the reliance dataset.

We compared the F-score for the humans' decisions between the Pred-RC condition and random condition, which was provided from the reliance dataset. Here, the F-score was calculated with the number of times human decisions matched the AI's success.

Hypothesis

We hypothesized that by properly selecting the timing for providing RCCs, Pred-RC can soften the decrease in the F-score in spite of the RCCs omitted, whereas that of the random condition was larger.

Results

Figure 5 illustrates the F-score for the humans' decisions. We conducted an ANCOVA[4] to statistically analyze the results. There were significant effects for the number of cues

[3]Confidence interval

[4]Analysis of covariance

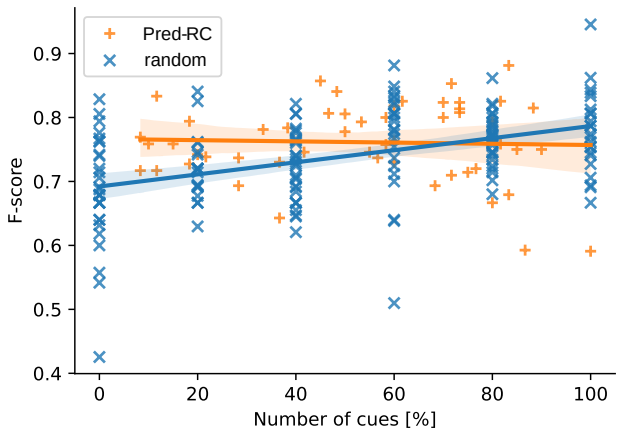


Figure 5: F-score of humans' decisions. Error bands shows 95% CI for linear regressions.

i	1	2	3	4	5	6
x						
$y^{\mathrm{AI}} = y^*$	true	true	true	true	false	false
c	0.89	[MASK]	[MASK]	[MASK]	0.09	0.21
d	AI	AI	AI	AI	human	human
decision correct?	✓	✓	✓	✓	✓	✓

Figure 6: Successful example of Pred-RC

($F(1,199) = 30.1; p < .0001; \eta_p^2 = .132$), the RCC selection method ($F(1,199) = 4.54; p = .034; \eta_p^2 = .022$), and the interaction effect ($F(1,199) = 7.31; p = .007; \eta_p^2 = .035$).

In the random condition, the F1-score decreased as the number of RCCs decreased. On the other hand, as far as the range of the data, the number of RCCs had little effect on the F-score in the Pred-RC condition, so the difference in F1-score between the Pred-RC and random conditions broadened as RCCs were reduced. This suggests that Pred-RC softened the decrease in the F-score in spite of the omitted RCCs, which supports our hypothesis. Therefore, we conclude that Pred-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.

Discussion

Examples of Pred-RC's behavior

Figure 6 shows a successful example of Pred-RC. We need to mention that it is difficult to follow the actual dynamics of the interaction among Pred-RC, participants, and tasks, so our explanations here are post-hoc. First, four images from the dataset on which the task AI achieved the best accuracy were provided, and actually, all the answers for them by the AI were correct. Pred-RC decided to show a high confidence rate at the first image, and the participant could correctly assign it to the AI. Pred-RC did not provide an RCC for the next three

i	1	2	3	4	5	6
x						
$y^{AI} = y^*$	false	false	false	false	true	true
c	[MASK]	[MASK]	[MASK]	[MASK]	[MASK]	0.63
d	AI	AI	AI	human	human	AI
decision correct?	✗	✗	✗	✓	✗	✓

Figure 7: Unsuccessful example of Pred-RC

images, which we can consider a valid decision because they were similar to the first one, so the participant was likely to expect the AI to be continuously successful. As a result, the participant could properly assign them to the AI. The fifth and sixth images were not in the datasets with which the task AI was trained, so the participant needed to do it by him/herself. Here, Pred-RC provided RCCs to avoid over-reliance on the AI, and the participant could avoid assigning them to the AI.

Figure 7 illustrates an unsuccessful example. The participant chose the AI for the first three tasks, observed the AI's failures in a row, and eventually changed his/her decision for the fourth task. Here, Pred-RC could not provide RCCs until the sixth task because the target percentage of RCCs was 20%, and *threshold* was high. At the fifth task, we found that the reliance model predicted a high reliance rate for both with and without an RCC (96.2% and 95.6%, respectively), which we can consider a false prediction because the participant had observed the failures and was likely to engage in under-reliance. A possible reason for this result is that the reliance model overfitted the visual features of the dataset of images with a blue background color. Since the task AI's accuracy for the dataset was high, and the participants tended to notice this, they also tended to assign images from this dataset to the AI. The model may have overestimated human reliance and been less sensitive to history data when images from this dataset were given. In contrast, the model and Pred-RC made a reasonable prediction and decision at the sixth task. The model predicted a high reliance rate with the RCC (65.7%) and low without it (34.9%). Because the success probability was 85.4%, Pred-RC decided to provide the RCC, and the participant could successfully choose the AI.

Participants' comments

We asked the participants to freely comment about their experiences during the CC task and acquired comments from 45 of them.

Twenty-one participants mentioned that they had focused on specific visual features such as "blue background color" or "dots" 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.

Six participants mentioned the provided RCCs. While one negatively assessed them ("I felt the confidence rate was not reliable"), the others provided positive comments ("I found that the AI was likely to succeed when the confidence rate was more than 50%, so I chose the AI then."). This may indicate that expectations toward the reliability of RCCs are differ by the individual, which may arouse distrust in RCCs. Using meta-RCCs, which calibrate trust not in an AI but RCCs, or multiple RCCs to make up for distrusted ones is promising as well. Pred-RC can theoretically afford multiple RCCs for its input by changing c_i, which is a future direction for extending Pred-RC. Depending on the situation, the problem of distrusted RCCs should be handled in another way such as apologies, excuses, or explanations and dialogues, which are found to be effective for trust repair (Lucas et al., 2018; Natarajan & Gombolay, 2020; Robinette, Howard, & Wagner, 2015; Sebo, Krishnamurthi, & Scassellati, 2019; Strohkorb Sebo, Traeger, Jung, & Scassellati, 2018).

Limitation

An important limitation of Pred-RC is that it 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. In the CC task, humans were not perfect as well (84.4% accuracy when $d_i =$ human). While our experiments successfully demonstrated that Pred-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.

We attempted to control the number of RCCs by changing *threshold*. However, in actual use, 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 the threshold. A possible approach is using reinforcement learning (RL), in which another reliance model learns not r_i but *threshold* with a reward function that balances the performance and cost.

Conclusion

This paper proposed Pred-RC, a method for selectively providing RCCs for human-AI collaboration. It decides whether to provide an RCC to avoid a discrepancy between the task success probability of an AI and the human reliance rate. In Pred-RC, a cognitive reliance model is used to predict reliance on an AI given a specific task on the basis of the collaboration history with a human. It can predict whether a human will assign a task to an AI for cases in which an RCC is provided or not. We implemented and tested Pred-RC for a human-AI collaboration task called the CC task. The results demonstrated that Pred-RC can perform reliance calibration with a reduced number of RCCs, indicating that we can selectively provide RCCs by evaluating their influence on human reliance with reliance prediction.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *6*, 52138-52160. doi: 10.1109/ACCESS.2018.2870052

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2020, jan). Trust-aware decision making for human-robot collaboration: Model learning and planning. *J. Hum.-Robot Interact.*, *9*(2). Retrieved from https://doi.org/10.1145/3359616 doi: 10.1145/3359616

Choo, S., & Nam, C. S. (2022). Detecting human trust calibration in automation: A convolutional neural network approach. *IEEE Transactions on Human-Machine Systems*, *52*(4), 774-783. doi: 10.1109/THMS.2021.3137015

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In R. Shumaker & S. Lackey (Eds.), *Virtual, augmented and mixed reality. designing and developing virtual and augmented environments* (pp. 251–262). Cham: Springer International Publishing.

Dubois, C., & Ny, J. L. (2020). Adaptive task allocation in human-machine teams with trust and workload cognitive models. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 3241-3246.

Fett, A.-K. J., Gromann, P. M., Giampietro, V., Shergill, S. S., & Krabbendam, L. (2012, 11). Default distrust? An fMRI investigation of the neural development of trust and cooperation. *Social Cognitive and Affective Neuroscience*, *9*(4), 395-402. Retrieved from https://doi.org/10.1093/scan/nss144 doi: 10.1093/scan/nss144

Fukuchi, Y., Osawa, M., Yamakawa, H., & Imai, M. (2022). Explaining intelligent agent's future motion on basis of vocabulary learning with human goal inference. *IEEE Access*, *10*, 54336-54347. doi: 10.1109/ACCESS.2022.3176104

Gul, F., Rahiman, W., & Alhady, S. S. N. (2019). A comprehensive study for robot navigation techniques. *Cogent Engineering*, *6*(1), 1632046. Retrieved from https://doi.org/10.1080/23311916.2019.1632046 doi: 10.1080/23311916.2019.1632046

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th international conference on machine learning - volume 70* (p. 1321-1330). JMLR.org.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., ... Tao, D. (2022). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-1. doi: 10.1109/TPAMI.2022.3152247

Häuslschmid, R., von Bülow, M., Pfleging, B., & Butz, A. (2017). Supportingtrust in autonomous driving. In *Proceedings of the 22nd international conference on intelligent user interfaces* (p. 319-329). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3025171.3025198 doi: 10.1145/3025171.3025198

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013). Presenting system uncertainty in automotive uis for supporting trust calibration in autonomous driving. In *Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications* (p. 210-217). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2516540.2516554 doi: 10.1145/2516540.2516554

Hoff, K., & Bashir, M. (2015, May 23). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. doi: 10.1177/0018720814547570

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53-71. doi: 10.1207/S15327566IJCE0401_04

Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., ... Levine, S. (2018, 29–31 Oct). Scalable deep reinforcement learning for vision-based robotic manipulation. In A. Billard, A. Dragan, J. Peters, & J. Morimoto (Eds.), *Proceedings of the 2nd conference on robot learning* (Vol. 87, pp. 651–673). PMLR.

Krber, M., Baseler, E., & Bengler, K. (2018, 01). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18-31. doi: 10.1016/j.apergo.2017.07.006

Lucas, G. M., Boberg, J., Traum, D., Artstein, R., Gratch, J., Gainer, A., ... Nakano, M. (2018). Getting to know each other: The role of social dialogue in recovery from errors in social robots. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (p. 344-351). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3171221.3171258 doi: 10.1145/3171221.3171258

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. In *Proceedings of the 11 th australasian conference on information systems* (pp. 6–8).

Matsumori, S., Shingyouchi, K., Abe, Y., Fukuchi, Y., Sugiura, K., & Imai, M. (2021, October). Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In *Proceedings of the ieee/cvf international conference on computer vision (iccv)* (p. 1898-1907).

McGuirl, J., & Sarter, N. (2006, 02). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, *48*, 656-65. doi: 10.1518/001872006779166334

Mousavi, S., Ellsworth, W., Weiqiang, Z., Chuang, L., & Beroza, G. (2020, 08). Earthquake transformer-an attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*, 3952. doi:

10.1038/s41467-020-17591-w

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction* (p. 33-42). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3319502.3374839 doi: 10.1145/3319502.3374839

Okamura, K., & Yamada, S. (2020a, 02). Adaptive trust calibration for human-ai collaboration. *PLOS ONE*, *15*(2), 1-20. Retrieved from https://doi.org/10.1371/journal.pone.0229132 doi: 10.1371/journal.pone.0229132

Okamura, K., & Yamada, S. (2020b). Empirical evaluations of framework for adaptive trust calibration in human-ai cooperation. *IEEE Access*, *8*, 220335-220351. doi: 10.1109/ACCESS.2020.3042556

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253. Retrieved from https://doi.org/10.1518/001872097778543886 doi: 10.1518/001872097778543886

Rai, A. (2020). Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141.

Ren, F., & Bao, Y. (2020). A review on human-computer interaction and intelligent robots. *International Journal of Information Technology & Decision Making*, *19*(01), 5-47. Retrieved from https://doi.org/10.1142/S0219622019300052 doi: 10.1142/S0219622019300052

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. In A. Tapus, E. André, J.-C. Martin, F. Ferland, & M. Ammi (Eds.), *Social robotics* (pp. 574–583). Cham: Springer International Publishing.

Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *2016 11th acm/ieee international conference on human-robot interaction (hri)* (p. 101-108). doi: 10.1109/HRI.2016.7451740

Scharowski, N., Perrig, S. A. C., von Felten, N., & Brühlmann, F. (2022). Trust and reliance in xai – distinguishing between attitudinal and behavioral measures. In *Chi 2022 workshop on trust and reliance in ai-human teams.* Retrieved from https://arxiv.org/abs/2203.12318 doi: 10.48550/ARXIV.2203.12318

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). "i don't believe you": Investigating the effects of robot trust violation and repair. In *Proceedings of the 14th acm/ieee international conference on human-robot interaction* (p. 57-65). IEEE Press.

Sheng, S., Pakdamanian, E., Han, K., Wang, Z., Lenneman, J., & Feng, L. (2021). Trust-based route planning for automated vehicles. In *Proceedings of the acm/ieee 12th international conference on cyber-physical systems* (p. 1-10). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3450267.3450529 doi: 10.1145/3450267.3450529

Strohkorb Sebo, S., Traeger, M., Jung, M., & Scassellati, B. (2018). The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 acm/ieee international conference on human-robot interaction* (p. 178-186). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3171221.3171275 doi: 10.1145/3171221.3171275

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

von Ahn, L., Blum, M., Hopper, N. J., & Langford, J. (2003). Captcha: Using hard ai problems for security. In E. Biham (Ed.), *Advances in cryptology — eurocrypt 2003* (pp. 294–311). Berlin, Heidelberg: Springer Berlin Heidelberg.

Walker, F., Verwey, W., & Martens, M. (2018, 06). Gaze behaviour as a measure of trust in automated vehicles..

Yagoda, R. E., & Gillan, D. J. (2012). You want me to trust a robot? the development of a human-robot interaction trust scale. *International Journal of Social Robotics*, *4*, 235-248.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (p. 295-305). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3351095.3372852 doi: 10.1145/3351095.3372852