IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 27, 2020, accepted November 20, 2020, date of publication December 4, 2020,
date of current version December 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3042556

Empirical Evaluations of Framework for Adaptive Trust Calibration in Human-AI Cooperation

KAZUO OKAMURA[1] **AND SEIJI YAMADA**[1,2]**, (Member, IEEE)**
[1]Department of Informatics, The Graduate University for Advanced Studies, SOKENDAI, Tokyo 101-8430, Japan
[2]National Institute of Informatics, Tokyo 101-8430, Japan

Corresponding author: Kazuo Okamura (kazu.ms+ia@gmail.com)

ABSTRACT Recent advances in AI technologies are dramatically changing the world and impacting our daily life. However, human users still essentially need to cooperate with AI systems to complete tasks as such technologies are never perfect. For optimal performance and safety in human-AI cooperation, human users must appropriately adjust their level of trust to the actual reliability of AI systems. Poorly calibrated trust can be a major cause of serious issues with safety and efficiency. Previous works on trust calibration have emphasized the importance of system transparency for avoiding trust miscalibration. Measuring and influencing trust are still challenging issues; consequently, not many studies have focused on how to detect improper trust calibration nor how to mitigate it. We approach these research challenges with a behavior-based approach to capture the status of calibration. A framework of adaptive trust calibration is proposed, including a formal definition of improper trust calibration called ''a trust equation''. It involves cognitive cues called ''trust calibration cues (TCCs)'' and a conceptual entity called ''trust calibration AI'' (TCAI), which supervises the status of trust calibration. We conducted empirical evaluations using a simulated drone environment with two types of cooperative tasks: a visual search task and a real-time navigation task. We designed trust changing scenarios and evaluated our framework. The results demonstrated that adaptively presenting a TCC could promote trust calibration more effectively than a traditional system transparency approach.

INDEX TERMS Trust, trust calibration, human-AI cooperation, trusted AI, human-agent interaction.

I. INTRODUCTION

AI technologies have become increasingly common in all aspects of our life. Examples of application areas include autonomous vehicles, medical services, virtual agents, and various web services. In such applications, it is inevitable that human users will need to cooperate appropriately with AI systems as such technologies are never perfect. One key aspect of human-AI cooperation is that human users should trust AI systems, just as humans normally do with other human partners [1], [2]. Trust definitely impacts human behavior and the outcome of cooperation [3]–[5]. Trust is an attitudinal judgment of the degree to which a user can rely on an agent to achieve their goals under conditions of uncertainty [6].

Successful cooperation between users and agents would require the users to appropriately adjust their level of trust to the actual reliability of AI systems. This process is called ''trust calibration'' [2]. While the reliability of an AI system changes for various reasons in an environment, users often fail to calibrate their trust in an AI system and end up in a status called ''over-trust'' or ''under-trust.'' Over-trust is poorly calibrated trust that exceeds the reliability of an AI system; it can result in over-reliance on an AI system with the expectation that it can perform outside of its designed capability. Over-trust sometimes leads to serious safety problems such as accidents involving autonomous vehicles [7], [8]. Under-trust is poorly calibrated trust that falls short of the AI's reliability; it can result in an agent not being used, excessive user workload, or deterioration in the total system performance [9].

To help over-trusting or under-trusting users re-calibrate their trust, we need to measure trust and influence it. However, these two elements are still challenging.

Measuring trust is not easy, as trust is a latent construct. Most of the research on trust has used self-reported trust scales [10]–[12]; however, they are so intrusive that it is

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

not practical to use them during task execution. Trust questionnaires conducted at the end of an experiment sometimes do not correctly reflect real-time trust during the experiment [13]. Some studies examined the effectiveness of physiological and neural measures such as gaze, heart rate, and EEG. Although these are promising approaches, further research would be necessary to clarify the correlation between trust and these metrics. Several studies explored not measuring trust but estimating it by using formal models. Nam *et al.* [14] proposed a trust model formulated as a Markov decision process based on the physical characteristics of robot swarms. Chen *et al.* [15] modeled human trust as a latent variable in a partially observable Markov decision process that could accommodate trust dynamics and human decision models. The trust modeling approach deeply depends on the task properties or specific behaviors of robots. Azevedo-Sa *et al.* [16] proposed an estimation method which integrates driver's behavior data through a Kalman filter-based approach.

Managing trust by manipulating factors proven to be influential in developing trust would also be complicated and difficult. Extensive research has been done examining the factors influencing trust or antecedents of trust. The goal of such research is to capture the most critical variables that might have causal links to human trust [17]–[19]. Hoff and Bashir [20] reported 29 factors that are influential in the development of human trust. Schaefer *et al.* [3] listed 31 factors. In both studies, they demonstrated that there are many interactions among these factors and showed that some of them are context-dependent or specific to human characteristics. Although these findings are significantly valuable in analyzing the latent structures of human trust, they also suggest that it would be difficult to influence human trust intentionally just by manipulating these factors.

Not many studies have focused on how to detect improper trust calibration nor how to mitigate it. This paper aims to address this deficiency in existing literature. We approach the research challenges with an emphasis on two important aspects of trust in human-AI cooperation: performance and human behavior. We previously proposed a method of adaptive trust calibration [21], using a formal definition of over-trust and under-trust, and conducted an initial evaluation with an over-trust scenario. In the current study, we extend the original method by introducing a third actor called "trust calibration AI" (TCAI) to human-AI cooperation. TCAI, which was originally discussed in [22], is a meta-level conceptual entity that supervises the status of trust calibration, to human-AI cooperation. This allows us to more clearly define the process of supporting trust calibration. The contributions of this paper are as follows.

- We have proposed a framework of adaptive trust calibration by extending the original method with a conceptual entity called "TCAI."
- The results of two empirical evaluations demonstrate that the proposed framework was successful in detecting and mitigating trust miscalibration that occurred in two different types of cooperative applications: a visual search and real-time navigation.
- We have discussed the applicability of the proposed framework by classifying the types of human-AI cooperation.

The remainder of the current paper is organized as follows. Section II reviews the existing work on trust calibration, our proposed framework is explained in Section III, section IV gives an overview of the empirical evaluations, section V describes the first evaluation with a visual search task, section VI describes the second evaluation with a real-time navigation task, and general discussions and the conclusion are provided in Section VII and VIII.

II. RELATED WORK

Many attempts have been made to evaluate the effects of system transparency in keeping appropriate trust. For an automated decision support system, McGuirl *et al.* [23] showed that presenting continually updated system-confidence information could improve trust calibration and lead to better performance in a human-machine team. Studies on visualizing a car's level of uncertainty during autonomous driving [24]–[26] have indicated that good transparency by presenting system information helps maintain the appropriate trust in vehicles. Helldin *et al.* [24] did experiments with 59 drivers in a simulated autonomous driving environment. They demonstrated that the drivers of autonomous vehicles who were provided with uncertainty information trusted the automated system less than those who did not receive such information, which indicates more proper trust calibration than in the control group. The drivers with the uncertainty information also took control of the car faster when needed and were able to perform tasks other than driving without risking safety.

The primary goal of realizing system transparency is to avoid improper trust calibration, not to deal with the status of over-trust and under-trust.

Recent studies on trust repair in human-robot interaction can be viewed as one of the countermeasures against under-trust situations. Marinaccio *et al.* [27] proposed a framework for repairing trust based on four types of errors that can occur when automated aids are used in a healthcare system: slips, lapses, mistakes, and violations. Their framework provides effective repair strategies according to the type of error. Robinette *et al.* [28] examined whether robots can repair trust by apologizing, promising to do better, and providing additional information relevant to the trust situation. They showed that all three of these actions can work and are more effective when robots use them just prior to human users deciding to trust them. Liu *et al.* [29] proposed a trust-repairing method for human-supervised teams of robot swarms. They proposed an algorithm for correcting undesired swarm behaviors by weighting the information shared among the robots in the swarm on the basis of the difference between the desired goal and the current behavior of the robots. The results of an online

experiment showed that their proposed method was effective at restoring trust. Tolmeijer *et al.* [30] developed a taxonomy of potential trust violations and suitable repair strategies for human-robot interaction. Their taxonomy was defined with four failure types and nine mitigation strategies. On the basis of the taxonomy, conceptual ideas of autonomous failure detection and repair were presented, using techniques such as formal verification, explanation, and trust loss detection. They suggested that a formal logic for trust [31] could be used to model robotic trust scenarios to identify when and how a system is not trusted or trust is lost.

As shown above, the existing studies on trust repair focused on the functionalities of robots for restoring trust. In contrast, we propose a framework for encouraging humans to re-calibrate their trust by detecting the miscalibration status of either over-trust or under-trust and emitting a simple cue to inform them of it.

III. PROPOSED FRAMEWORK

We propose a framework of adaptive trust calibration that consists of three elements:

1) definitions of over-trust and under-trust called ''trust equations'',
2) a cognitive cue called a ''trust calibration cue'',
3) a conceptual entity called a ''trust calibration AI''.

To describe the framework, we first define human-AI cooperation as a series of actions taken by a human user *repeatedly working on selection problems* to decide on either AI execution or manual execution. Both the human user and the AI should have the same functionality to execute a common task, with different levels of performance depending on the situation. The human user must solve a problem by selecting who is to execute the task, and the final responsibility for the outcome always belongs to the human user.

A. TRUST EQUATIONS

Performance and human behaviors in human-AI cooperation play critical roles in our definitions of over-trust and under-trust called ''trust equations''. Achieving better performance is one of the fundamental goals of human-AI cooperation. Previous research showed that trust in robots is mainly affected by a robot's performance [32]. Therefore, we focus on the performance-related factors that influence trust. This focus makes it possible to narrow down the definition of trust to ''the expectation that a task done by an AI system will be successful.'' The estimated reliability of an AI system in terms of performance can be a good index of such an expectation. Trust can also be viewed as a human user's behavior [33] in choosing whether to rely on an AI system or to do a task manually. From a performance point of view, such observable choice behavior can be considered a result of comparing the estimated reliabilities of humans and AI.

Three performance-related parameters, P_A, $\widehat{P}_A$, and P_H, are defined as follows.

- P_A: Probability that a task done by an AI system will be successful. This is called the ''reliability of the AI system.''
- $\widehat{P}_A$: Human user's estimation of P_A. This is a *user's trust in the AI system*.
- P_H: Probability that a task done manually by a human user will be successful. This is called the ''capability of the user.''

The reliability of the AI system P_A varies depending on the conditions of the AI system. The user's trust $\widehat{P}_A$ also changes accordingly and becomes equal to P_A if trust is appropriately calibrated. Over-trust occurs if $\widehat{P}_A > P_A$, and under-trust occurs if $\widehat{P}_A < P_A$. Since measuring the user's trust $\widehat{P}_A$ is difficult, we modified the definitions of over-trust and under-trust by using a third parameter P_H in addition to $\widehat{P}_A$ and P_A:

- **Over-trust**: the human user estimates that the AI system is more reliable than the user even though the actual reliability of the AI system is lower than the user's capability.

$$(\widehat{P}_A > P_H) \wedge (P_H > P_A) \tag{1}$$

- **Under-trust**: the user estimates that they are better at a task than the AI system even though the actual reliability of the system is higher than the user's capability.

$$(\widehat{P}_A < P_H) \wedge (P_H < P_A) \tag{2}$$

We call these two definitions ''trust equations''. The first terms of (1) and (2) can be calculated by observing the user's behaviors. Several studies [34]–[36] have demonstrated that reliance behavior can be explained by the relationship between a user's trust in a system and the user's self-confidence in performing a task manually. Maehigashi *et al.* [37] found that human users select for a task to be done either through automation or manually on the basis of how they perceive their own manual performance. When a user decides to use a system, it is reasonable to say that this behavior indicates $\widehat{P}_A > P_H$. If the user chooses a manual execution, it indicates $\widehat{P}_A < P_H$. The first terms, which are the inequalities of $\widehat{P}_A$ and P_H, can be judged by observing the user's behavior, without directly measuring $\widehat{P}_A$ or P_H. If the second terms can be estimated, we can identify the trust calibration status of human-AI cooperation using the trust equations.

B. TRUST CALIBRATION CUE

To effectively notify human users of improper trust calibration, we explore the idea of giving them simple cues when over-trust or under-trust is detected. Once users fall into the state of over-trust or under-trust, it might not be easy to get out of the state. This may be an example of confirmation bias in the preservation of trust [38]. Calibration occurs only in response to new evidence that changes the users' situational awareness, while new evidence cannot be learned without changing the current behavior first [33]. To solve this dilemma, a new trigger is necessary to make the user aware of

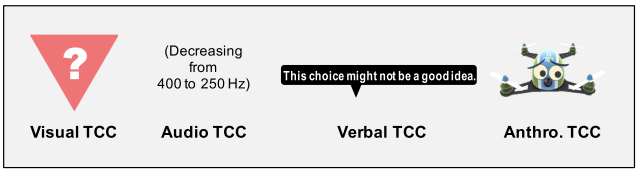

FIGURE 1. Four types of TCCs.

changes in the environment and to re-calibrate trust. We call this trigger the "trust calibration cue" (TCC).

In our previous work [21], we designed and evaluated four different types of TCCs: visual, audio, verbal, and anthropomorphic (see Figure 1). The visual TCC was a red warning sign, which is considered to be one of the most common alert signs according to [39], [40]. The audio TCC was a sound with a frequency that decreases from 400 Hz to 250 Hz as a negative message [41]. The verbal TCC is a tooltip balloon with the warning message "This choice might not be a good idea." The anthropomorphic TCC was an animated drone image with a cartoon-like face. The results of the evaluation indicated that the verbal TCC was most effective among the four TCCs in changing human users' reliance behaviors.

The concept of the TCC was inspired by the works done by Komatsu *et al.* [41], who proposed an intuitive notification method called "artificial subtle expressions" (ASE). One of their design requirements is "complementary," which means that notifications should not interfere with the main communication protocol. A TCC should also be perceived through a different channel than the one used for the main interface of a Task-AI. Users with improper trust calibration are likely to have difficulty in understanding the system information coming thorough the main interface.

Several studies on trust have proposed the use of "cues" to increase system transparency. Visser *et al.* [42] proposed a design guideline for trust cues, which are informational elements for assessing trust with autonomous systems. They defined the cues in terms of trust dimensions and trust processing stages. Cai and Lin [43] examined multi-modal cues for conveying the confidence of a driver assistance system. Unlike our TCC, the purpose of these "cues" was to deliver system information to users.

C. TRUST CALIBRATION AI

Figure 2 shows a diagram of the proposed framework for adaptive trust calibration. We introduce a conceptual entity in human-AI cooperation called "trust calibration AI" (TCAI), which is a meta-level entity that manages the whole process of adaptive trust calibration. The TCAI is invisible to the user to avoid bringing about issues regarding trust in the TCAI.

The human user works on selection problems to decide whether a task should be done by an AI system (called a "Task-AI" in this diagram) or the human user should do it manually. The Task-AI provides a human user with system information through its system transparency interface. The human user makes decisions by comparing P_H and $\widehat{P}_A$, which are his/her reliability and the estimated reliability of the Task-AI. Each decision corresponds to the first inequalities in the proposed framework, (1) and (2).

The TCAI observes the human's choice behaviors, which indicate the answers to selection problems. This observation is made to evaluate the first inequalities in the trust equations. The system transparency interface of the Task-AI, which discloses its internal information, can help human users solve selection problems. The TCAI also solves the selection problems by estimating P_A and $\widehat{P}_H$ with a model-based or statistical approach. These estimations correspond to evaluating the second inequalities in the trust equations. If the observed human behaviors are not consistent with the TCAI's estimations, the TCAI judges that it has detected over-trust or under-trust according to the trust equations, and it gives a TCC to the human user to notify the user of an improper trust calibration status. Although the TCAI can solve the selection problems, it is always the human user, not the TCAI nor the Task-AI, who makes the final decisions since the human user is fully responsible for the outcomes of human-AI cooperation. The TCAI only suggests to the human user to recalibrate trust in the Task-AI.

The basic algorithm of the adaptive trust calibration performed by the TCAI is described in Algorithm A0. This framework aims to adaptively prompt a user to calibrate her/his trust by presenting a trust calibration cue only when the TCAI detects over-trust or under-trust by observing the user's choice behavior. This approach is taken to mitigate over-trust or under-trust, in contrast with the traditional approach of trying to maintain appropriate trust calibration with continuous system transparency.

Algorithm A0 Adaptive Trust Calibration by TCAI

while Cooperative tasks exist **do**
  Observe a user's choice behavior. $\cdots \widehat{P}_A \lessgtr P_H$
  Evaluate the second inequalities
   of the trust equations (1) and (2). $\cdots \widehat{P}_H \lessgtr P_A$
  Detect improper trust calibration.
  if over-trust or under-trust is detected **then**
    Present a trust calibration cue to the user.
  end if
end while

IV. OVERVIEW OF TWO EMPIRICAL EVALUATIONS

We conducted two empirical studies to evaluate whether the proposed framework was effective for two different types of collaborative tasks. The evaluations were done with online experiments using a web-based 3D drone simulator, which we developed based on an open-source JavaScript WebGL library CesiumJS [44] and the Bing Map API [45]. A screenshot of the simulator running on a Chrome browser is shown in Figure 3.[1]

[1]All map images in the current paper are from Geospatial Information Authority of Japan (CC BY 4.0). The images are similar but not identical to the original ones used in the experiments due to a copyright reason.

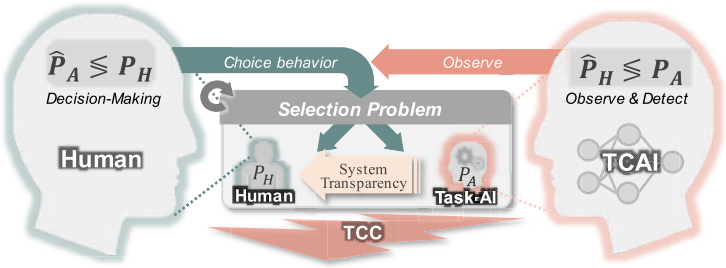

FIGURE 2. Proposed framework for adaptive trust calibration.

FIGURE 3. Online drone simulator. Simple operations with two cursor keys for controlling the drone and mouse buttons for making choices.

The first empirical study was done with **pothole inspection**, which is a series of visual search tasks to check if there are any holes or cracks in road images from a drone. Each inspection task is executed *discretely* automatically by the Task-AI or manually by a human user, and the tasks are *mutually independent*. This type of visual inspection can be categorized as a reconnaissance task, which is often used in the trust research literature.

The second empirical study was done with **autonomous drone navigation**. which is a series of real-time control tasks of navigating a drone to reach a goal along a predefined course. Each task is *continuously* executed through cooperative activities between auto-pilot, operated by the Task-AI, and manual-pilot, done by a human user. The state of each navigation task is *dependent* on the result of the previous navigation task, in terms of the drone position and the direction. Autonomous driving (SAE level 4) falls into the same category as this task.

We expected the participants of the experiments to change their choice behaviors if the TCAI detected inappropriate trust calibration and presented a TCC. We tested the following hypotheses:

- **[H1]** the choice rates of manual executions increase if TCCs are presented in cases of over-trust or decrease if TCCs are presented in cases of under-trust.
- **[H2]** the participants with TCCs perform better than the participants without TCCs.
- **[H3]** adaptively presenting TCCs could trigger the trust calibration process more effectively than continuously maintaining system transparency.

Participants were recruited through crowdsourcing services. Regarding online experiments in general, Crump *et al.* [46] showed that the data collected online using a web-browser seemed mostly in line with laboratory results, so long as the experiment methods were solid. In these two online experiments, sound effects and simulator speed performances might vary with the different PC environments used by each participant. To minimize these differences, 1) we instructed the participants to adjust the audio volume properly before the experiments, and 2) the simulator performance was automatically measured and adjusted to achieve the same drone speed at each participant's PC. We also checked IP addresses to avoid duplicate participation by the same person.

Both empirical studies were carried out with written informed consent from all participants in accordance with the Declaration of Helsinki and the recommendations of the Ethical Guidelines for Medical and Health Research Involving Human Subjects provided by the Ministry of Education, Culture, Sports, Science and Technology and the Ministry of Health, Labour and Welfare in Japan. The protocol was approved by the ethics committee of the National Institute of Informatics.

V. EVALUATION WITH POTHOLE INSPECTION

This section presents the first empirical study to evaluate the proposed framework with pothole inspection tasks under the bi-directional changes of trust conditions. An early report on this evaluation was described in [47].

A. METHOD

1) POTHOLE INSPECTION TASKS

A route with 30 checkpoints (CKPs) was prepared. Each CKP was located in a rectangular inspection area. CKPs on the route were displayed as small yellow circles on the screen and as a pink one if it was the next target. The ten CKPs had potholes in the corresponding areas while the other twenty did not. When the drone came close enough to the next target CKP on the route, a message popped up (Figure 4) in which the participants were asked to make a choice: automatic inspection or manual inspection.

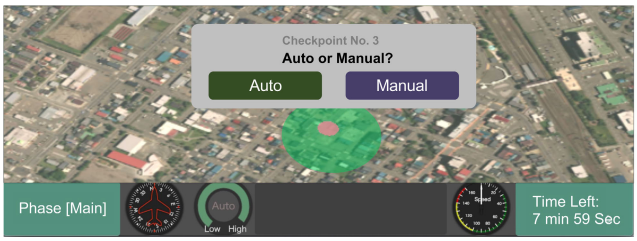

FIGURE 4. Popup message asking the participants for choice.

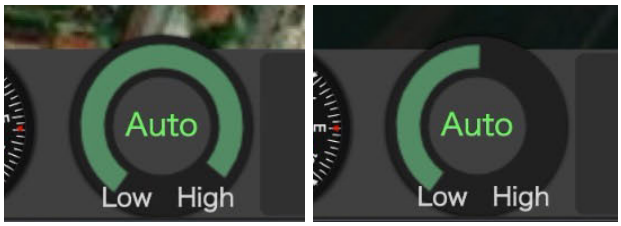

FIGURE 5. Reliability Indicator at the bottom left area of the screen. Showing a higher reliability (left) and a deteriorated reliability (right).

FIGURE 6. Pothole inspection. Automatic inspection result window (left). Manual inspection window (right).

The reliability of the automatic inspection was always displayed by the indicator at the bottom left area of the screen (Figure 5), which increased the system transparency.

If the participants clicked the "Auto" button, an automatic-inspection result with a road image of the inspected area was shown for three seconds. This feedback information helped the participants evaluate the automatic inspection performance, thereby increasing the system transparency [2], [48]. If the participants selected the "Manual" button, a road image was popped up, and the participants had to make a pothole report manually. Figure 6 shows the popup windows of both cases. Potholes were artificially shown as irregular shapes in a dark brown color on a road image in the popup window.

We used a verbal TCC in this experiment as it showed the most significant effect to change users' behaviors in the first empirical study. The screen image of the verbal TCC is shown in Figure 7. If the proposed framework detected over-trust or under-trust from a participant choice, this TCC was presented right after the choice action(pushing a button).

FIGURE 7. Verbal TCC.

2) PARTICIPANTS AND SCENARIOS

A total of seventy participants (51 male, 19 female) took part in the experiment online. Their ages ranged from 25 to 75 years old ($M = 44.2$, $SD = 10.3$). The participants were recruited through a cloud-sourcing service provided by Yahoo! Japan.

We defined the ABA/BAB scenarios of under-trust (A) and over-trust (B) by manipulating the weather conditions. The performance of the automatic pothole inspection P_A was configured on the basis of signal detection theory (SDT) [49]. SDT defines the detection of signals in noisy environments. Noise and signals are represented as overlapping density distributions. The distance between the two distributions represents the sensitivity d' of a system. In the A condition, the weather conditions were set to be good in the simulated environment, and P_A and the corresponding sensitivity d' were manipulated to be 0.88 and 2.35, respectively, indicating that the agent has a very high discrimination ability. In contrast, the weather conditions were bad in the B condition, and P_A dropped to 0.50, and the corresponding sensitivity d' became 0.1, indicating the greatly deteriorated reliability of the automatic pothole inspection.

If the participants failed to calibrate their trust properly, the possibility of under-trust in the A condition or over-trust in the B condition would be higher. In the ABA scenario, the weather conditions of the experiment started as A, then changed to B, and finally went back to A. The same applies to the BAB scenario. Each condition continued until eight CKPs were inspected. Participants were randomly assigned to one of four groups: the NoTCC-ABA group (without TCC in the ABA scenario), TCC-ABA (with a verbal cue in the ABA scenario) group, NoTCC-BAB group, and TCC-BAB group. The NoTCC-ABA/BAB groups were control groups in this experiment.

3) PROCEDURES

The online experiment started with an **instruction phase** (see Figure 8). The participants were given an instruction stating that the goal of the experiment was to inspect 24 CKPs within 20 minutes. They also learned that the average success rate of manual pothole inspection was 75%. The drone's automatic inspection was explained as "The reliability is almost perfect, close to 100%," for the participants of the two groups in the ABA scenario and "The automatic inspection

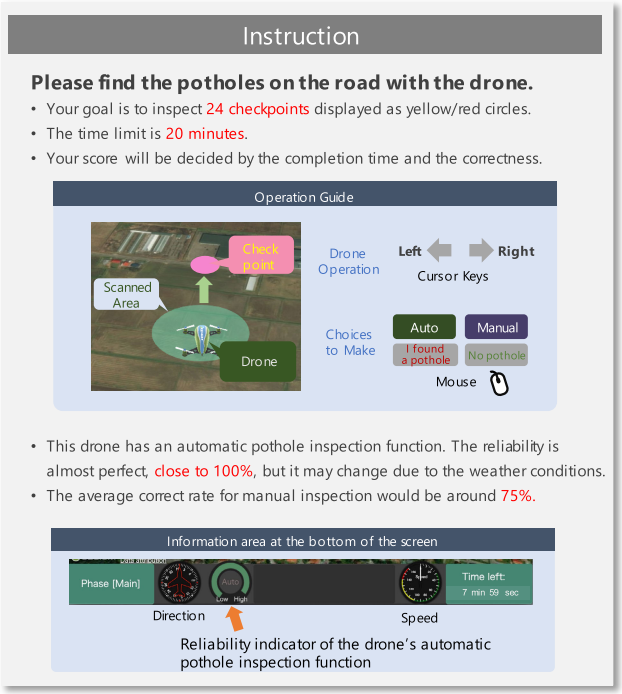

FIGURE 8. The instruction screen.

is accurate" for the participants of the two groups in the BAB scenario. These sentences were meant to help the participants calibrate their initial trust properly in the first period. They also understood that the reliability of the automatic inspection could fluctuate depending on the weather conditions. This instruction was given to help the participants calibrate their trust properly when the condition changed. In this instruction phase, the participants were also guided to adjust the sound volume level by listening to a 400-Hz beep sound.

Next, in the **training phase**, the participants started a practice flight of the drone and learned how to inspect the CKPs. This phase was finished after the first three CKPs were inspected, and **the main phase** of the experiment was started. The main phase first started with either condition A or B depending on the scenario of the group. In the A condition, the weather was good, and the visibility in the simulated environment was high. Therefore, the drone's automatic inspection functioned very well. In the B condition, it was dark and rainy with the sound effects of a thunderstorm.

The reliability of the automatic inspection deteriorated due to the low visibility in the environment. Each condition continued until the participants completed the inspection of eight CKPs. The 1st CKP, the 9th CKP, and the 17th CKP were the first CKPs of the three conditions. Figure 9 illustrates the manipulation of P_A with the weather conditions and the expected changes of P_H.

If the participants completed the 24th inspection or the elapsed time exceeded 20 minutes, the main phase of the experiment was finished.

The algorithm A1 "Adaptive Trust Calibration (1)" based on the proposed framework was applied in the experiment.

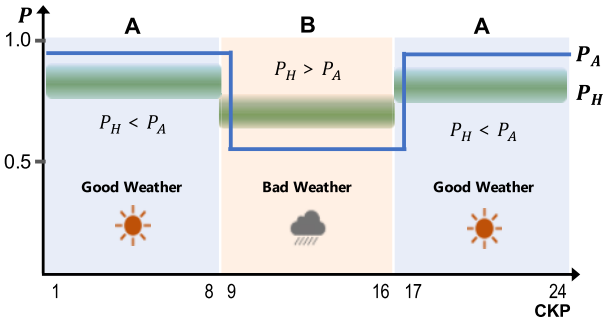

FIGURE 9. ABA scenario.

Algorithm A1 Adaptive Trust Calibration (1)

Initialize:
M = the total number of CKPs.;
Over-trust flag list: OT[1], ..., OT[M] are initialized with zero;
Under-trust flag list: UT[1], ..., UT[M] are initialized with zero;
The number of current CKP: $i \Leftarrow 1$;

while $i \leqq M$ **and** not time-over **do**
  if the drone reached a CKP **then**
    if choice behavior is AUTO **and** $P_H > P_A$ **then**
      $OT[i] \Leftarrow 1$;
      if $i \geqq 3$ *and* $(OT[i-2] + OT[i-1]) \geqq 1$ **then**
        Over-trust is detected and TCC is presented to the user;
      end if
    else if choice behavior is MANUAL **and** $P_H < P_A$ **then**
      $OU[i] \Leftarrow 1$;
      if $i \geqq 3$ *and* $(OU[i-2] + OU[i-1]) \geqq 1$ **then**
        Under-trust is detected and TCC is presented to the user;
      end if
    end if
    $i \Leftarrow i + 1$;
  end if
end while

A simple moving average of three CKPs was used in the algorithm to capture the participants' behavior changes in each condition with eight CKPs.

4) ESTIMATION OF P_H AND MANIPULATION CHECK

Geirhos *et al.* [50] demonstrated that human image recognition is still better than the top-performing deep neural networks in the case of image degradation such as Gaussian blur or additive Gaussian noise. As the pothole inspection tasks are mainly image recognition tasks with blurred and noisy road images when the weather conditions turned worse, we assumed that P_H would not fluctuate more widely than P_A when the weather conditions changed. Thus the inequality

$P_A > P_H$ was estimated to be true during the good weather period and false during the bad one.

As a pre-experiment, we measured the manual success rates (P_H) with the prepared CKP data to verify our assumption. Thirty-two participants (25 male, 7 female) were recruited through a cloud-sourcing service provided by Yahoo! Japan. Their ages ranged from 25 to 65 years old ($M = 42, SD = 12$). None of them joined the main experiment. They manually inspected the prepared CKPs following the same procedure of the main experiment, except that there was no automatic inspection available. Half of them were in the A condition, and the other were in the B condition. The results indicated that the mean of the manual success rates and the sensitivity d' was 0.83 ($SD = 0.15$) and 1.85 for the A condition and 0.79 ($SD = 0.15$) and 1.69 for the B condition. As already explained, the performance of the automatic inspection in the main experiment was manipulated so that the success rates and the sensitivity d' were 0.88 and 2.35 for condition A and 0.50 and 0.00 for condition B. One sample t-test showed that the manual success rate was smaller than the automatic success rate for the A condition [$t(47) = -2.26, p = 0.01, Cohen'sd = 0.33$] and larger than the automatic success rate for the B condition [$t(47) = -13.66, p < 0.01, Cohen'sd = 1.97$]. Therefore, we concluded that our assumption on P_H was valid with the prepared CKP data for the main experiment.

5) THE DEPENDENT VARIABLES

In this experiment, TCC presentation rates (hereinafter called "TCC rates"), manual choice rates (hereinafter called "manual rates"), and the sensitivity d' were measured as the dependent variables. TCC rates are the rates of the frequency at which TCCs were presented to the participants at each CKP, indicating how our framework was working during the experiment. Manual rates are the mean values of the manual choice ratio for each condition, showing how the participants relied (or did not rely) on the drone's automatic inspection and therefore indicating their trust status. The sensitivity d' demonstrates the performance of human-AI collaborative tasks.

B. RESULTS

Seventy participants completed all 24 CKPs within the time limit. Of the seventy participants, 17 were in the NoTCC-ABA group, 18 in the TCC-ABA group, 21 in the NoTCC-BAB group, and 14 in the TCC-BAB group. The average time taken to finish the main phase of the experiment was 9 minutes 5 seconds, which means 22.5 seconds per CKP.

1) TCC RATES

Figure 10 and Figure 11 illustrate the TCC rates at each CKP of the TCC-ABA group and the TCC-BAB group. Table 1 shows 3-CKP means of TCC rates in each condition. C1, C2, and C3 mean A, B, and A for the ABA groups, B, A, and B for the BAB groups. 'i-j' indicates the mean value of the TCC rates from CKP i to CKP j. Standard errors are in parentheses.

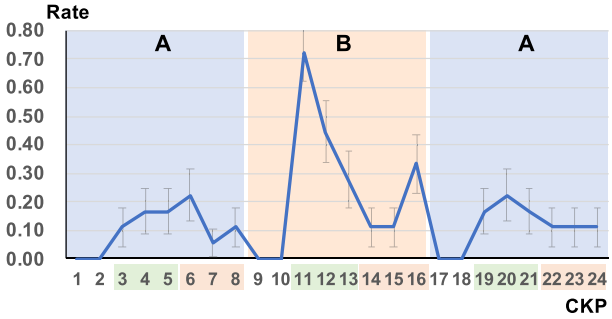

FIGURE 10. TCC rates of TCC-ABA group.

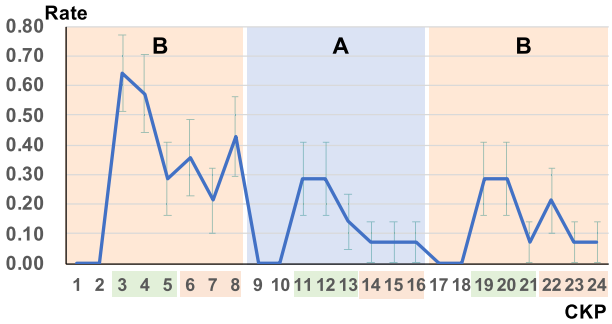

FIGURE 11. TCC rates of TCC-BAB group.

TABLE 1. 3-CKP means of TCC rates in each condition. Standard errors are in parentheses.

	C1		C2		C3	
CKP i-j	3-5	6-8	11-13	14-16	19-21	22-24
TCC-ABA	0.15 (0.08)	0.13 (0.07)	0.48 (0.11)	0.19 (0.08)	0.19 (0.08)	0.11 (0.07)
TCC-BAB	0.5 (0.13)	0.33 (0.13)	0.24 (0.11)	0.07 (0.07)	0.21 (0.11)	0.12 (0.09)

ABA groups: The mean of the TCC rates from CKP 3 to 5 [hereinafter referred to as MR (3-5)] for the first A condition (C1) was low at 0.15 and slightly decreased to 0.13 for MR (6-8). We did a paired t-test that revealed no significant difference between MR (3-5) and MR (6-8). For the B condition (C2), the TCC rate went up to the maximum at CKP 11. MR (11-13) was 0.48 and quickly decreased after that. A paired t-test showed that MR (14-16) was significantly lower than MR (11-13) [$t(17) = 4.53$, one-tailed, $p<0.01$, Cohen's $d = 0.99$]. For the second A condition (C3), the TCC rates were almost the same as the first A condition. The difference between MR (19-21) and MR (22-24) was not statistically significant.

BAB groups: The TCC rate for the first B condition (C1) started from the highest value among the conditions at CKP 3 and then decreased with some fluctuations. A paired t-test

TABLE 2. Means of the manual rates and the sensitivity d'.

	Manual rate			Sensitivity d'		
Condition	C1	C2	C3	C1	C2	C3
NoTCC-ABA	0.23 (0.08)	0.28 (0.09)	0.26 (0.07)	1.67 (0.05)	1.12 (0.14)	1.74 (0.10)
TCC-ABA	0.19 (0.06)	0.50 (0.06)	0.22 (0.07)	1.46 (0.12)	1.25 (0.10)	1.80 (0.04)
NoTCC-BAB	0.46 (0.08)	0.32 (0.08)	0.63 (0.09)	0.53 (0.21)	1.39 (0.12)	0.67 (0.26)
TCC-BAB	0.45 (0.09)	0.22 (0.08)	0.71 (0.06)	0.88 (0.20)	1.47 (0.10)	0.73 (0.21)

revealed that MR (6-8) was significantly lower than MR (3-5) [t(13) = 1.84, one-tailed, p = 0.04, Cohen's d = 0.43]. For the A condition (C2), a paired t-test revealed that MR (14-16) significantly decreased from MR (11-13) [t(13) = 1.99, one-tailed, p = 0.03, Cohen's d = 0.53] in the TCC-BAB group. For the second B condition (C3), there was no significant difference between MR (19-21) and MR (22-24), although TCC rates slightly decreased during the condition.

In summary, the TCC rates for all conditions showed a similar trend in which the values were initially higher and then decreased along the CKP series, except for the first A condition of the TCC-ABA group. Higher TCC rates were observed for the B conditions than the A conditions. This indicates that over-trust detections in the bad weather were more frequent than under-trust detections in the good weather.

2) MANUAL RATES

The change in manual rates indicates how the participants changed their trust in the automatic inspection. Building trust is an accumulating process [6], and TCCs might need some time to have an effect on changing manual rates and also might be presented more than once per participant. Therefore, we evaluated the proposed framework by comparing the eight-CKP mean values of the manual rates for each condition so that we could capture the accumulated effects of presenting TCCs. Table 2 shows the means of the manual rates and the sensitivity d' for each condition. C1, C2, and C3 are either condition A or B, depending on the groups. Standard errors are in parentheses. We conducted a one-way ANOVA (within-subjects design; independent variable: the scenario conditions of three levels, A, B, and A (B, A, and B), dependent variable: manual rate) for each group. All post-hoc analysis was done using the Holm-Bonferroni method. Figure 12 illustrates the manual rates for each condition of each groups.

ABA groups: The result of the ANOVA for the NoTCC-ABA group did not show any significant difference in the manual rates among the three conditions [$F(2, 32) = 0.20$, $p = 0.82$, $\eta_p^2 = 0.01$]. In comparison, the ANOVA for the TCC-ABA group revealed a significant difference in the manual rates among the conditions in the ABA scenario [$F(2, 34) = 6.50$, $p < 0.01$, $\eta_p^2 = 0.28$]. The post-hoc

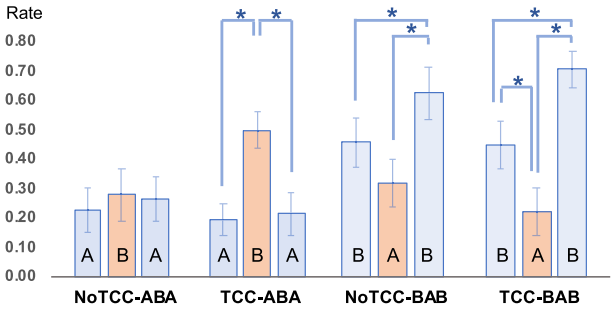

FIGURE 12. Manual rates ($*$: $p < 0.05$).

analysis indicated that the manual rate for the B condition significantly increased from the first A condition [$t(17) = 3.56$, $adjusted.p < 0.01$]. The manual rate for the second A condition also significantly decreased [$t(17) = 2.45$, $adjusted.p = 0.03$] from the B condition, and the manual rates for the first A condition and second A condition were not significantly different [$t(17) = 0.79$, $adjusted.p = 0.79$].

BAB groups: The ANOVA analysis for the NoTCC-BAB group revealed that there was a significant difference in the manual rates [$F(2, 40) = 6.41$, $p < 0.01$, $\eta_p^2 = 0.24$]. The post-hoc analysis showed that the rate for the B condition was not significantly changed from that for the first A condition [$t(20) = 1.46$, $adjusted.p = 0.16$], while the manual rate for the second B condition significantly increased from the A condition [$t(20) = 3.14$, $adjusted.p = 0.02$], and it was also significantly larger than for the first B condition [$t(20) = 2.84$, $adjusted.p = 0.02$]. The ANOVA analysis for the TCC-BAB group showed that there was a significant difference in the manual rate [$F(2, 26) = 14.48$, $p < 0.01$, $\eta_p^2 = 0.53$]. The post-hoc analysis indicated that the manual rate for the A condition significantly decreased from the first B condition [$t(13) = 2.65$, $adjusted.p = 0.02$]. For the second B condition, the manual rate increased significantly from the A condition [$t(20) = 4.47$, $adjusted.p < 0.01$].

3) PERFORMANCE

We conducted the same one-way ANOVA with the sensitivity d' of each group. Figure 13 illustrates the sensitivity d' for each condition of each groups.

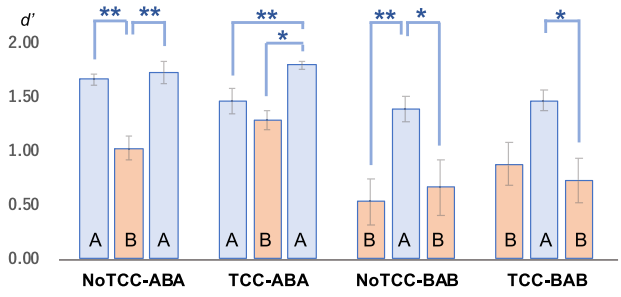

FIGURE 13. Sensitivity d' ($*: p < 0.05, **: p < 0.01$).

ABA groups: For the NoTCC-ABA group, the main effect of the sensitivity d' was found to be significant [$F(2, 32) = 14.8, p < 0.01, \eta_p^2 = 0.48$]. The post-hoc analysis indicated that the mean value of d' significantly decreased from the first A condition to the B condition [$t(16) = 5.26$, $adjusted.p < 0.01$] and then significantly increased from the B condition to the second A condition [$t(16) = 4.05$, $adjusted.p < 0.01$]. For the TCC-ABA group, the main effect of the sensitivity d' was found to be significant [$F(2, 34) = 7.52, p < 0.01, \eta_p^2 = 0.31$]. The post-hoc analysis indicated that the mean value of d' significantly increased from the B condition to the second A condition [$t(17) = 5.44$, $adjusted.p < 0.01$] and also showed a significant increase from the first A condition to the second A condition [$t(17) = 2.61$, $adjusted.p = 0.04$].

BAB groups: For the NoTCC-BAB group, the main effect of the sensitivity d' was found to be significant [$F(2, 40) = 7.45, p < 0.01, \eta_p^2 = 0.27$]. The post-hoc analysis revealed that the mean value of d' significantly increased from the first B condition to the A condition [$t(20) = 3.76$, $adjusted.p < 0.01$] and significantly decreased from the A condition to the second B condition [$t(20) = 2.98$, $adjusted.p = 0.01$]. For the TCC-BAB group, the main effect of the sensitivity d' was found to be significant [$F(2, 26) = 4.75, P = 0.02, \eta_p^2 = 0.27$]. The post-hoc analysis indicated that the mean value of d' for the A condition marginally increased from that for the first B condition [$t(13) = 2.46$, $adjusted.p = 0.06$]. The mean value of d' for the second B condition significantly decreased from that for the A condition [$t(13) = 3.13$, $adjusted.p = 0.02$].

C. DISCUSSION

1) ABA SCENARIO

For **the first A condition**, the TCC rates of both the NoTCC-ABA group and the TCC-ABA group were low. The manual rates were also low for both. This suggests that the participants in both groups properly calibrated their trust in the high reliability of the automatic inspection under the good weather conditions, probably on the basis of their knowledge acquired in the initial instruction phase. For **the B condition**, the status of trust in the previous condition was clearly carried over, so the TCC rates were initially very high. This suggests that most of the participants were initially over-trusting the drone's automatic capability even when its reliability became very low under the bad weather conditions. The TCC rates drastically dropped for the TCC-ABA group. The manual rates significantly increased for this group, while that for the NoTCC-group remained the same as in the previous condition. The sensitivity d' for the B condition was kept high for the TCC-ABA group, while that for the NoTCC-ABA group significantly dropped under the bad weather conditions. These results indicate that presenting TCCs in the B condition greatly impacted how participants behaved in making choices, and the results also suggest that they could properly calibrate their trust. Consequently, their task performance did not deteriorate despite the bad weather. For **the second A condition**, the manual rates of the TCC-ABA group significantly decreased from the previous condition, while those of the NoTCC-ABA group did not change at all. It is not explicitly clear whether the participants in the NoTCC groups properly calibrated their trust for this condition; however, the task performance of the NoTCC groups was slightly worse than that in the TCC-ABA group.

BAB scenario: For **the first B condition**, the TCC rates were high at the beginning. This was probably caused by the instruction given to the participants regarding the high reliability of the automatic inspection. After the initial high period, the TCC rates showed a statistically significant decrease for this condition. Although the mean values of manual rates both for the NoTCC-BAB group and the TCC-BAB group were almost similar in this condition, the sensitivity d' indicates that the TCC-BAB group performed better than the NoTCC-BAB group. For **the A condition**, the TCC rates started at a slightly higher level than those observed for the other A conditions in the ABA scenario. The rates steadily decreased and reached the lowest levels among all conditions in the experiment. The manual rates of the TCC-BAB group showed a statistically significant drop from the previous condition, while that of the NoTCC-BAB group did not. The performance of the TCC-BAB group was kept higher than that of the NoTCC-BAB group. These results demonstrate the effectiveness of presenting TCCs to affect the behaviors of the participants for whom the status of trust was undertrust and suggest that trust calibration done to mitigate undertrust was successfully promoted by the proposed framework. For **the second B condition**, the TCC rates decreased toward the end of this condition with some fluctuations. The manual rates of both groups significantly increased to the highest values in the experiment. One possible interpretation would be that the 16 tasks before the second B condition would be enough for most of the participants to learn the system and the environment so that the participants in the NoTCC-BAB group could calibrate their trust better in the second B condition. Similar learning effects might also be behind the low manual rates in the second A condition of the ABA scenario.

The TCC groups significantly changed their choice behaviors over the first two conditions both in the ABA and in the BAB scenarios, while the TCC groups did not. These results clearly support hypothesis **H1**. Regarding the performance, the results of the sensitivity d' confirm hypothesis **H2**, except

for the case that the mean value of d' for the A condition of the TCC-ABA group was slightly smaller than that of the NoTCC-ABA group of which the participants probably calibrated the trust properly.

The weather changes from the A condition to the B condition or vice versa were very noticeable in terms of screen visibility and sound effects. Nevertheless, the participants of the NoTCC-ABA group did not significantly change their choice behaviors and they were over-trusting or under-trusting the drone's automatic inspection. The reliability information continuously displayed at the reliability indicator did not help the participants to calibrate the trust properly. In contrast to this, the participants in the TCC groups successfully altered their choice behaviors at the first weather changes. We believe that the results demonstrate the effectiveness of the adaptive method and confirmed hypothesis **H3**. TCCs were given immediately after the behavior only if the TCAI judged the participants to be in a state of over-trust or under-trust, so it would be easier for them to understand the message of the cues and to move forward in the trust calibration process.

Although we observed the under-trust status in the A condition of the BAB scenario, the over-trust status was more obviously observed in the B condition of the ABA scenario. One of the reasons would be that the instruction of the experiment made the participants expect the higher reliability of the automatic inspection. Existing studies also demonstrated the human tendency toward the automation called automation bias [23] or perfect automation schema [51].

VI. EVALUATION WITH AUTONOMOUS DRONE NAVIGATION

This section presents the second empirical study to evaluate the proposed framework in a real-time application environment. An early report on this evaluation was described in our previous work [52]. We designed a cooperative control task of navigating a drone to reach a destination along a predefined course. The navigation can be done either by the drone's automatic capability or by a manual control. In contrast to the pothole inspection tasks used in the first empirical evaluation, the participants' selection decisions and operations must be made quickly enough to control the drone smoothly.

A. METHOD

1) AUTONOMOUS DRONE NAVIGATION TASKS

We added an auto-pilot function to the drone simulator used in the previous experiments. Figure 14 shows a new screen image of the simulator running in the Chrome browser.

The participants performed a task in which they flew a drone along a course that was displayed on a screen until the drone reached the goal of the course. A 10-km course was prepared with an average altitude of 214 meters. The course consisted of three 3.3-km parts (see Figure 15) with the same trajectory in terms of curve and height. The width of the course was 10.4 meters. The participants had to control the drone so that it stayed on the course until the goal. The drone could be flown by autonomous navigation. This type of control is called "auto-pilot." The auto-pilot was implemented with a PID control over the heading direction and the pitch of the drone to minimize cross-track error (see Figure 16), which is the shortest distance between the drone and the center line of the prepared course. The reliability of the auto-pilot was always shown on the indicator displayed at the bottom area of the screen.

The participants could take over the navigation of the drone at any time with the left or right cursor keys. This control is hereinafter called "manual-pilot." The manual-pilot period expired after 1.5 seconds unless any further key inputs occurred. In this experiment, the pitch control was always under auto-pilot, and the roll of the drone was fixed flat to make the manual-pilot easier. The level of automation in the experiment corresponded to Level 4 of the autonomous driving [53], meaning that the auto-pilot could fly the drone at all times, and participants could take over the control if they wanted to, but they were not required to do so.

FIGURE 14. Online semi-autonomous drone simulator.

FIGURE 15. The first part of the course.

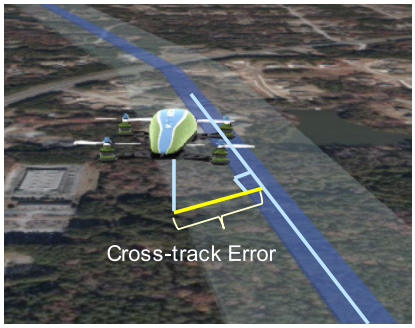

FIGURE 16. Cross-track error.

FIGURE 17. Verbal TCC.

The verbal TCC was presented in front of the drone (see Figure 17) when over-trust and under-trust were detected by the framework. The message was intentionally indirect so as to encourage the participants to re-consider their decisions rather than blindly follow a cue.

To apply the proposed framework to the autonomous drone navigation tasks, we modified the behavior measurement to capture users' choice behaviors.

Let b_i be a sampled behavior at a timing i ($0 \leq i \leq N$), where $b_i = \{1 : \text{reliance}, 0 : \text{no reliance}\}$, and N is the maximum sampled timing of the task. Let B_t be a moving average of b_i at a timing t ($t \geq W$).

$$B_t = \frac{1}{W} \sum_{i=t-W}^{t} b_i, \tag{3}$$

where W ($0 \leq w \leq N$) is the size of the time window defined in accordance with the characteristics of the cooperation task. Let K be a specified threshold. If $B_t > K$, it means $\widehat{P}_A > P_H$. Otherwise, it indicates $\widehat{P}_A < P_H$.

The second terms of (1) and (2), P_A could be calculated with the sensor models and algorithms used to implement the system, and P_H could be estimated by using the parameters of a target task and environmental conditions. Therefore, the second terms can be also estimated.

2) PARTICIPANTS AND SCENARIO

A total of 36 online participants (30 male, 6 female) were recruited a cloud-sourcing service provided by Yahoo! Japan. Participants were randomly assigned to one of two groups: the NoTCC group (without TCC) and the TCC group (with TCC). Four of the male participants failed to complete the experiment due to large deviations from the course. This left us 32 participants whose ages ranged from 22 to 70 years old (M = 46.6, SD = 11.4).

We defined the ABA scenario of under-trust (A) and over-trust (B) by manipulating the reliability of the auto-pilot. In the two A conditions, good weather conditions were simulated. The screen brightness was 100%, and there were no sound effects except for the sound of the drone flying. The parameters of the PID control were configured so that P_A became 0.93 and 0.91 in the first A (A1) condition and the second A (A2) conditions respectively, which means the drone with auto-pilot flew accurately along the course. In the B condition, a thunderstorm was simulated with a blurred and dark (40% brightness) screen and with sound effects. The cross-track errors, which were inputs to the PID control, were artificially distorted to simulate the deteriorated sensing accuracy under the bad weather conditions. This made P_A deteriorate to 0.69, and the drone with auto-pilot would thus often be off course. The participants were expected to take over the control of the drone (called "disengagement" in autonomous driving) when they saw the drone with auto-pilot fail to stay on course.

3) PROCEDURES

The online experiment started with **an instruction phase**. The participants were given an instruction stating that the goal of the experiment was to fly the drone along the 10-km course within 15 minutes. They were told that the score would be better if the flight was more accurate. They learned that the reliability of the drone's auto-pilot, which was continuously displayed on the indicator, was very high, although it could fluctuate depending on the weather conditions. Next, **the training phase started**. The participants started a practice flight of the drone and experienced both the auto-pilot and the manual-pilot with some guidance on the screen. The speed of the drone was automatically adjusted according to the performance of the PC of each participant to equalize the conditions of the experiment. This phase was finished when the drone reached the end of the 3-km training course, and the main phase of the experiment was started with the A condition. The proposed detection framework was applied during this phase. The first A condition (hereinafter, called the A1 condition) changed to the B condition followed by the second A condition (hereinafter, called the A2 condition). Each condition lasted for 3.3 km. When the drone reached the goal of the 10-km main course or the elapsed time exceeded 15 minutes, the main phase was finished. After the experiment, the participants were asked to fill out a post-experiment questionnaire. The algorithm A2 "Adaptive Trust Calibration (2)" based on the proposed framework was applied in the experiment. The while loop in the algorithm was implemented as a timer-event handling loop in the experimental system. The timer-event was fired every 0.12 second. The moving average of the participants behavior were calculated at each timer-event. The window size was 12 seconds ($W = 100$), which was suitable for capturing the changes in trajectory for the prepared course.

4) ESTIMATION OF P_H AND MANIPULATION CHECK

We assumed that the drone's auto pilot would utilize a visual SLAM algorithm in the real situations to locate its position. Although the robust algorithms are proposed, low-illumination scenes remain challenging tasks [54]. Moreover, as described before, the work of [50] demonstrated that human image recognition is still better than the top-performing deep neural networks in the case of image degradation. These pieces of work could provide a basis for estimating the second terms of the proposed framework in the experiment. We assumed that P_A would fluctuate more widely than P_H under changing weather conditions, and we estimated that the inequality $P_A > P_H$ was true during the good weather period and false during the bad one. We did a pre-experiment to measure P_H by asking the participants

Algorithm A2 Adaptive Trust Calibration (2)

Initialize:
$W \Leftarrow 100$; $K \Leftarrow 0.5$:
while the drone is not reached the goal **and** not time-over **do**
  Get *SampledBehavior*; /* 1:Auto or 0:Manual */
  Estimate P_H and P_A;
  if $MovingAve(SampledBehavior, W) > K$ **then**
    $Behavior \Leftarrow AutoPilot$;
  else
    $Behavior \Leftarrow ManualPilot$;
  end if
  if $Behavior = AutoPilot$ *and* $P_H > P_A$ **then**
    Over-trust is detected and TCC is presented to the user;
  else if $Behavior = ManualPilot$ *and* $P_H < P_A$ **then**
    Under-trust is detected and TCC is presented to the user;
  end if
end while

to fly the drone with the manual-pilot only. Twenty participants [17 male, 3 female, mean age 40.0 (SD = 12.0)] were recruited through a cloud-sourcing service provided by Yahoo! Japan. They performed the manual navigation tasks in accordance with the same procedure of the main experiment. The results indicated that the mean of the success rates of the manual-pilot were 0.79, 0.80, 0.81 for the A1 condition, the B condition and the A2 condition, respectively. One-sample t-tests revealed that $P_A > P_H$ in the A1 condition [$t(19) = -3.04, p < 0.01, Cohen'sd = 0.68$] and also in the A2 condition [$t(19) = -2.19, p = 0.04, Cohen'sd = 0.52$]. Another one-sample t-test indicated that $P_A < P_H$ in the B condition [$t(19) = 2.31, p = 0.03, Cohen'sd = 0.49$]. These results indicated that our assumptions were valid in the current experiment.

5) DEPENDENT VARIABLES

In this experiment, three things were measured as the dependent variables. TCC rates are the rates of the frequency at which TCCs were presented to the participants, indicating how our framework was working during the experiment. Manual-pilot rates are the mean values of the manual-pilot ratio for each condition, showing how the participants relied (or did not rely) on the drone's auto-pilot and therefore indicating their trust calibration status. The means of cross-track errors indicates the task performances or how well the collaborative flight tasks between auto-pilot and manual-pilot were done.

B. RESULTS

Of the 32 participants, 17 were in the NoTCC group and 15 were in the TCC group. The average time taken to finish the main phase of the experiment was XX minuets YY seconds.

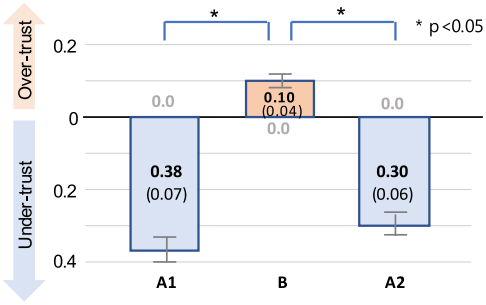

FIGURE 18. TCC rates.

1) TCC RATES

Figure18 illustrates TCC rates for each conditions in the TCC group. TCCs were presented when under-trust was detected in the A1 condition and A2 condition, and when over-trust was detected in the B condition. The result of a one-way ANOVA showed that the effect of the ABA conditions on the TCC rates was significant [$F(2, 28) = 6.41, p < 0.01, \eta_p^2 = 0.31$]. The post-hoc analysis using the Holm-Bonferroni method showed that TCC rates for the A1 condition was significant larger than that for the B condition [$t(14) = 2.77, adj.p = 0.045$] and the A2 condition [$t(14) = 2.72, adj.p = 0.045$]. TCC rates for the B condition was significant smaller than that for A2 condition [$t(14) = 2.17, adj.p = 0.048$].

2) MANUAL-PILOT RATES

Figure 19 shows the manual rates in each group. The result of the one-way ANOVA for the NoTCC group did not show any significant difference in the manual rates among the three conditions [$F(2, 32) = 1.60, p = 0.22, \eta_p^2 = 0.09$]. On the other hand, the one-way ANOVA for the TCC group revealed the effect of the conditions [$F(2, 28) = 32.6, p < 0.001, \eta_p^2 = 0.70$]. Post-hoc analysis using he Holm-Bonferroni method showed that the manual rates for the B condition was significantly larger than those for the A1 condition [$t(14) = 6.68, adj.p < 0.001$] and for the A2 condition [$t(14) = 5.72, adj.p < 0.001$].

3) CROSS-TRACK ERRORS

Figure 20 illustrates the mean values of the cross-track errors. To evaluate performances of collaborative tasks between auto-pilot and manual-pilot, the cross-track errors of the NoTCC group and the TCC groups are compared with that of the manual-pilot only group measured in the pre-experiment, and also with that of the auto-pilot only. Although the one-way ANOVA for the cross-track errors did not showed the significant difference among the groups, the multiple comparisons using the Holm-Bonferroni method indicated that the difference between the NoTCC group and the TCC group was close to significance [$t(25.3) = 2.49, adj.p = 0.06, Cohen's\ d = 2.77$].

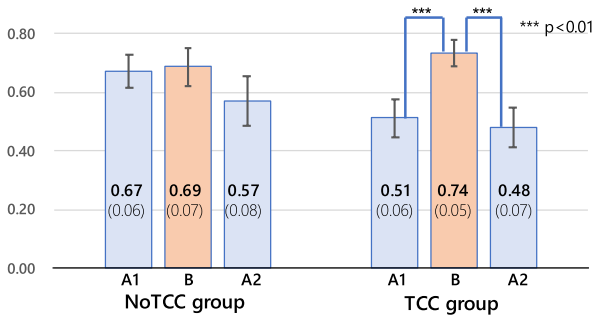

FIGURE 19. Manual-pilot rates.

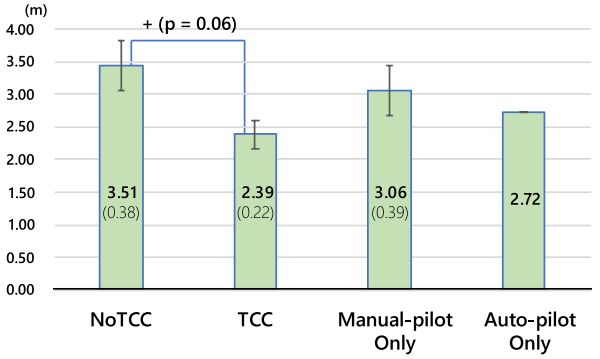

FIGURE 20. Cross-track errors.

C. DISCUSSION

In the A1 and A2 conditions, the TCC rates showed that the TCAI detected the under-trust and presented TCCs. The manual-pilot rates in the TCC group significantly increased from the A1 condition to the B condition, and significantly decreased from the B condition to A2 condition. No such changes among the conditions were observed in the NoTCC group. We consider these results supported the hypothesis **H1** in the case of under-trust. TCC groups showed smaller cross-track errors than the NoTCC group, although the difference was close to significance. Therefore, we consider that the results partially supported the hypothesis **H2**. The TCCs successfully triggered the participants to change their choice behaviors, and the task performances were improved. The manual-pilot rates in the NoTCC group were kept high even when the indicator showed high auto-pilot reliability in the A1 and A2 conditions. The system transparency provided by the indicator did not help the participants recalibrate their trust in the auto-pilot properly. In contrast, TCCs adaptively presented at the time of over-trust/under-trust changed the participants' behaviors in the TCC group. Therefore, these results supported the hypothesis **H3**.

In this experiment, the manual rates were higher than initially expected. This strong tendency of under-trust in both groups might be caused by the participants' preventive actions when they anticipated that drone was about to go out of the course. In the post-experiment questionnaire results, forty percent of the participants answered that they selected the manual-pilot when they notice that the drone was not heading along the course direction (even though the drone was still on the course). This early intervention observed in the experiment suggested that the drone's postures would significantly impact the participants' trust than the auto-pilot's reliability indicator.

Although we used a verbal TCC in this experiment, its calibration effect was milder than in the first evaluation. This result may be due to a mental workload of reading the text message of verbal TCC during the real-time navigation task. The other type of TCCs, such as the visual TCC or the audio TCC shown in Figure 1, could be more effective as they are more intuitive or with a different modality.

In summary, we demonstrated that our framework could promote trust calibration in the continuous real-time task. The task performance was also improved as a result of the proper trust calibration.

VII. GENERAL DISCUSSION

A. APPLICABILITY OF PROPOSED FRAMEWORK

The empirical studies described above demonstrated that the proposed framework helped the participants recalibrate their trust when they were performing the two types of cooperative tasks with the Task-AIs. In order to examine to what extent the proposed framework is applicable to various applications of human-AI cooperation, we discuss the prerequisites of the framework: types of cooperative sequences and performance-centric view of trust.

1) TYPES OF COOPERATIVE SEQUENCES

The trust equations are defined on the premise that human-AI cooperation is a series of actions taken by a human user repeatedly working on selection problems to decide on either AI execution or manual execution. Both the human user and the Task-AI can perform the task. Figure 21 illustrates the four possible sequences of task executions and selection decisions. The cooperative sequences discussed here are for a team that consists of one human user and one Task-AI to focus on the trust calibration issues that occur between them.

Types A0, A1, and A2 are within the scope of the proposed framework. The sequence shown in the upper left of this figure is Type A0. The human user should select who executes the task. The pothole inspection in the first evaluation corresponds to this type.

In Type A1, the Task-AI first executes a task, and the human user monitors the status. Then, the human user evaluates the result of the task execution and executes the task manually if necessary. Type A1 is a variation of Type A0 with the condition that the AI's task execution process and status can be monitored. The drone navigation in the second evaluation corresponds to this type. Applications of this type include SAE level 4 autonomous driving (driver and AI) [53], supervised unmanned vehicles (remote pilot and AI), and telepresence robot navigation (operator and AI).

If there is no resource competition required for the task execution and it is not necessary or possible for the human

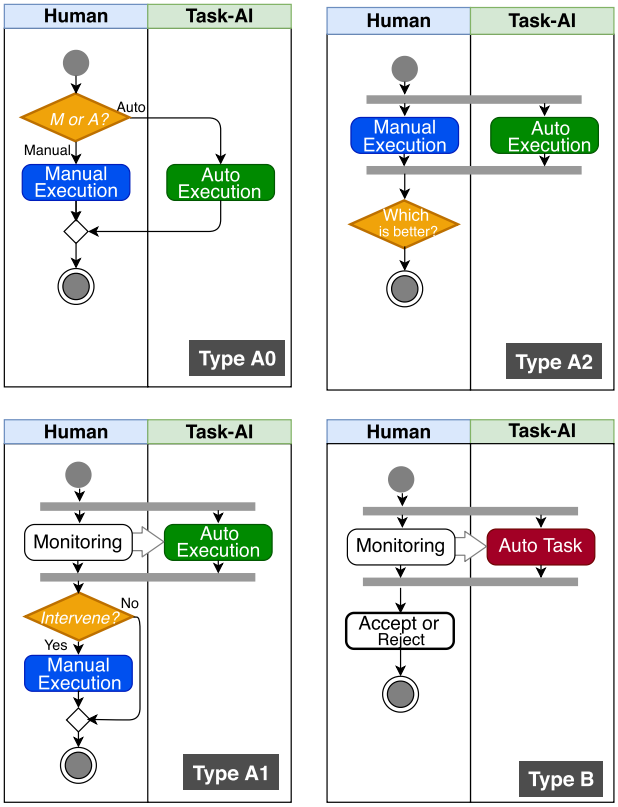

FIGURE 21. Types of cooperative sequences.

user to monitor the AI task execution, both the human and AI can execute the same task simultaneously, and the human can compare the results to choose which one to adopt. This sequence is illustrated as Type A2. Applications such as collaborative medical diagnostics (doctor and AI), product visual inspection (inspector and AI) and baggage screening systems (airport clerk and AI) could be examples of this type.

The cooperative sequence shown as Type B in Figure 21, where only AI can perform the target task, is beyond the scope of the current study. In Type B, the human user's tasks are monitoring the AI's task execution (if possible) and accepting or rejecting the result. Although it is clear that human trust in AI has a significant impact on the decision to accept or reject the result, the current framework cannot be applied to Type B applications where the human user is not capable of executing a task. Cooperative activities corresponding to this type include AI doctors (AI and patient) and SAE level 5 autonomous driving (AI and passenger).

According to the LOA (Levels of Automation) defined by Kaber and Endsley [55], [56], human-AI cooperation of Types A0, A1, and A2 would be at the level of "shared control," in which a human user decides on a particular option or strategy.

Although the cooperation sequences discussed here are in quite simple forms, there is a wide range of potential applications within the scope of the proposed framework.

2) PERFORMANCE-CENTRIC VIEW OF TRUST

We focus on trust factors related to system performance, as achieving higher performance is one of the most important goals of human-AI cooperation. If we can assume that a human user will act rationally and deterministically according to the estimated performance, trust can be viewed as the observable human behavior of selecting a better performance agent. In the proposed framework, the TCAI must be able to estimate two types of performance: the Task-AI's performance P_A and the human user's performance $\widehat{P}_A$. The TCAI knows how the Task-AI works, so it could use the internal system information to calculate the Task-AI performance, P_A. The human user's performance, $\widehat{P}_H$, could be estimated by the TCAI with a model-based or statistical approach. The results of the previous studies [114, 115] could provide a basis for such estimation. A top-down approach is considered a better way to build a model using prior knowledge about the cooperative task's features and structure. It is also useful to take a bottom-up approach that utilizes the data collected beforehand or on-the-fly during the task execution if an appropriate estimation model is not available.

B. ROLE OF TRUST CALIBRATION AI

Of the roles an AI system should play in human-AI cooperation, a Task-AI performs a domain-specific role and a TCAI performs a meta role of facilitating proper trust calibration.

Previous studies on human-agent teaming proposed their cooperation frameworks. Vecht *et al.* [57], [58] proposed a concept of social AI modules that serve as intelligent middleware aiming to transform task-oriented AI components and humans into a coherent human-agent team. One of the key functionalities of the social AI modules is to mediate high-level communication between humans and AIs. In their model, task-oriented AI components are designed to perform a specific task optimally but may not be optimized for human interaction. A pair of a task-oriented AI and a corresponding social AI module is equivalent to our Task-AI concept, which is designed to provide task-dependent information through its system transparency interface. Their model does not directly address trust calibration issues, which are the main target of our proposed framework with the TCAI. Cummings and Bruni [59] discussed three distinct roles in the cooperative decision-making process: the moderator, generator, and decider. The moderator in their process model is the agent that keeps the decision-making process moving forward. The generator is the agent that generates candidates of feasible solutions, and the decider is the agent that makes the final decision. In our proposed framework, which focuses on managing the trust calibration process in human-AI cooperation, the TCAI plays a similar role as the moderator and generator in their model. In addition to such functions, the TCAI encourages human users to recalibrate their trust by issuing TCCs so that the user could make better selections, thus achieving higher cooperation performances. The decider in our framework is always the human user.

The separation of a TCAI from a Task-AI allows the trust calibration process to be defined more clearly, therefore we expect that the proposed framework would facilitate a better design of collaborative systems.

VIII. CONCLUSION

We approached the research challenges in trust calibration with an emphasis on performance and human behavior in human-AI cooperation. In the current study, we extended our previously proposed method by introducing a conceptual entity called TCAI, which supervises the trust calibration process. We did two empirical studies to evaluate the proposed framework. The first evaluation revealed that the framework worked well with a visual search task involving pothole inspection under dynamic trust changes. The second evaluation indicated that the framework was also effective for trust calibration in a real-time control task involving drone navigation.

The recent proposal of Trust Engineering for human-AI teaming by Ezer *et al.* [60] insisted that there are still many challenges in managing trust in AI systems that are increasingly complex and work within imperfect information environments. They proposed six conceptual components in Trust Engineering: adaptability, communication, explainability, training/knowledge, assessment, and security. The results of the current study contribute to the first three components, which are mainly related to interactions between humans and AI.

Shneiderman [61] proposed a concept called human-centered artificial intelligence and discussed how to avoid the dangers of excessive human control or excessive computer control in Human-AI cooperation. In contrast to our framework, he emphasized human self-efficacy, mastery, and responsibility.

There are many other factors influencing trust to be considered in future research. Human-related factors such as personality, propensity to trust, or automation bias should be investigated further. The proposed detection algorithm made a binary decision with a simple moving average. Future research should involve exploring a different way of representing the over-trust or under-trust status, such as defining the status as a probability depending on the degree of miscalibration of trust. Although we have learned some lessons in the empirical studies indicating that the TCCs were more effective than a simple reliability indicator in the case of miscalibration, further research on multimodal user interfaces [62] would be necessary to evaluate the concept of TCCs.

The proposed framework for adaptive trust calibration has a simple structure that separates task-dependent and non-task-dependent parts, and it could be applied to many application situations. Despite several limitations, we believe that the framework could contribute to a baseline design of trustworthy systems for better human-AI cooperation.

REFERENCES

[1] D. M. Rousseau, S. B. Sitkin, R. S. Burt, and C. Camerer, "Not so different after all: A cross-discipline view of trust," *Acad. Manage. Rev.*, vol. 23, no. 3, pp. 393–404, Jul. 1998.

[2] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, Nov. 1994.

[3] K. E. Schaefer, J. Y. C. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Hum. Factors*, vol. 58, no. 3, pp. 377–400, May 2016.

[4] S. Lewandowsky, M. Mundy, and G. P. A. Tan, "The dynamics of trust: Comparing humans to automation.," *J. Experim. Psychol., Appl.*, vol. 6, no. 2, pp. 104–123, 2000.

[5] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, Mar. 1996.

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

[7] National Highway Traffic Safety Administration. (2017). *Automatic Vehicle Control Systems-Investigation of Tesla Accident*. [Online]. Available: https://static.nhtsa.gov/odi/inv/2016/INCLA-PE16007-7876.PDF

[8] National Highway Traffic Safety Administration. (2018). *Preliminary Report-Crash Involving Pedestria–Uber Test Vehicle*. [Online]. Available: https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf

[9] E. J. de Visser, M. M. Peeters, F. J. Malte, S. Kohn, H. S. Tyler, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *Int. J. Soc. Robot.*, vol. 12, no. 2, pp. 459–478, May 2020.

[10] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems," *Int. J. Cognit. Ergonom.*, vol. 4, no. 1, pp. 53–71, Mar. 2000.

[11] M. Madsen, "Measuring human-computer trust," in *Proc. 11th Australas. Conf. Inf. Syst.*, 2000, pp. 6–8.

[12] R. E. Yagoda and D. J. Gillan, "You want me to trust a ROBOT? The development of a human–robot interaction trust scale," *Int. J. Social Robot.*, vol. 4, no. 3, pp. 235–248, Aug. 2012.

[13] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *Proc. 8th ACM/IEEE Int. Conf. Hum.-Robot Interact. (HRI)*, Mar. 2013, pp. 251–258.

[14] C. Nam, P. Walker, M. Lewis, and K. Sycara, "Predicting trust in human control of swarms via inverse reinforcement learning," in *Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 528–533.

[15] 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. Hum.-Robot Interact.*, vol. 9, no. 2, pp. 1–23, Feb. 2020.

[16] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Real-time estimation of drivers' trust in automated driving systems," *Int. J. Social Robot.*, pp. 1–7, Sep. 2020.

[17] K. E. Oleson, D. R. Billings, V. Kocsis, J. Y. C. Chen, and P. A. Hancock, "Antecedents of trust in human-robot collaborations," in *Proc. IEEE Int. Multi-Disciplinary Conf. Cognit. Methods Situation Awareness Decis. Support (CogSIMA)*, Feb. 2011, pp. 175–178.

[18] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *Int. J. Man. Mach. Stud.*, vol. 27, pp. 527–539, Nov. 1987.

[19] J. K. Rempel, J. G. Holmes, and M. P. Zanna, "Trust in close relationships," *J. Pers. Soc. Psychol.*, vol. 49, no. 1, pp. 95–112, 1985.

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

[21] K. Okamura and S. Yamada, "Adaptive trust calibration for human-AI collaboration," *PLoS ONE*, vol. 15, no. 2, pp. 1–20, 2020.

[22] K. Okamura, "Adaptive trust calibration in human-AI cooperation," Ph.D. Dissertation, Dept. Inform., Graduate Univ. Adv. Studies (SOKENDAI), Hayama, Japan, 2020.

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

[24] 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. AutomotiveUI*, 2013, pp. 210–217.

[25] T. Helldin, "Transparency for future semi-automated systems," Doctoral Dissertation, Orebro Univ., Örebro, Sweden, 2014.

[26] R. Haeuslschmid, M. V. Buelow, B. Pfleging, and A. Butz, "Supporting trust in autonomous driving," in *Proc. 22nd Int. Conf. Intell. User Interfaces*, 2017, pp. 319–329.
[27] K. Marinaccio, S. Kohn, R. Parasuraman, and E. J. De Visser, "A framework for rebuilding trust in social automation across health-care domains," in *Proc. Int. Symp. Hum. Factors Ergon. Heal. Care*, vol. 4, no. 1, 2015, pp. 201–205.
[28] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *Proc. Int. Conf. Soc. Robot.*, 2015, pp. 574–583.
[29] R. Liu, Z. Cai, M. Lewis, J. Lyons, and K. Sycara, "Trust repair in human-swarm teams+," in *Proc. 28th IEEE Int. Conf. Robot Human Interact. Commun. (RO-MAN)*, Oct. 2019, pp. 1–6.
[30] S. Tolmeijer, A. Weiss, M. Hanheide, F. Lindner, T. M. Powers, C. Dixon, and M. L. Tielman, "Taxonomy of trust-relevant failures and mitigation strategies," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2020, pp. 3–12.
[31] A. Herzig, E. Lorini, J. F. Hubner, and L. Vercouter, "A logic of trust and reputation," *Log. J. IGPL*, vol. 18, no. 1, pp. 214–244, Feb. 2010.
[32] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Hum. Factors, J. Hum. Factors Ergonom. Soc.*, vol. 53, no. 5, pp. 517–527, Oct. 2011.
[33] B. D. Adams, L. E. Bruyn, and S. Houde, "Trust in automated systems literature review," Defence Res. Develop. Canada, Toronto, ON, Canada, Tech. Rep. CR2003-096, 2003.
[34] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *Int. J. Hum.-Comput. Stud.*, vol. 58, no. 6, pp. 719–735, Jun. 2003.
[35] J. Gao and J. D. Lee, "Extending the decision field theory to model operators' reliance on automation in supervisory control situations," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 36, no. 5, pp. 943–959, Sep. 2006.
[36] X. J. Yang, V. V. Unhelkar, K. Li, and J. A. Shah, "Evaluating effects of user experience and system transparency on trust in automation," in *Proc. ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2017, pp. 408–416.
[37] A. Maehigashi, K. Miwa, H. Terai, K. Kojima, and J. Morita, "Selection strategy of effort control: Allocation of function to manual operator or automation system," in *Proc. Annu. Meeting Cogn. Sci. Soc.*, 2011, pp. 1977–1982.
[38] D. Good, "Individuals, interpersonal relations, and trust," in *Trust: Making and Breaking Cooperative Relations*. Oxford, U.K.: Univ. Oxford, 2000, ch. 3, pp. 31–48.
[39] M. A. Changizi, M. Brucksch, R. Kotecha, K. McDonald, and K. Rio, "Ecological warnings," *Saf. Sci.*, vol. 61, pp. 36–42, Jan. 2014.
[40] K. R. Laughery and M. S. Wogalter, "A three-stage model summarizes product warning and environmental sign research," *Saf. Sci.*, vol. 61, pp. 3–10, Jan. 2014.
[41] T. Komatsu, S. Yamada, K. Kobayashi, K. Funakoshi, and M. Nakano, "Artificial subtle expressions: Intuitive notification methodology of artifacts," in *Proc. 28th Int. Conf. Hum. Factors Comput. Syst. (CHI)*, 2010, pp. 1941–1944.
[42] E. J. de Visser, M. Cohen, A. Freedy, and R. Parasuraman, "A design methodology for trust cue calibration in cognitive agents," in *Proc. Int. Conf. Virtual, Augmented Mixed Reality*, 2014, pp. 251–262.
[43] H. Cai and Y. Lin, "Tuning trust using cognitive cues for better human-machine collaboration," in *Proc. Hum. Factors Ergon. Soc.*, 2010, pp. 2437–2441.
[44] The Cesium Consortium. (2018). *CesiumJS-Geospatial 3D Mapping and Virtual Globe Platform*. [Online]. Available: http://cesiumjs.org
[45] Microsoft. (2018). *Bing Maps API Documentation*. [Online]. Available: https://www.microsoft.com/en-us/maps/
[46] M. J. C. Crump, J. V. Mcdonnell, and T. M. Gureckis, "Evaluating Amazon's mechanical turk as a tool for experimental behavioral research," *PLoS ONE*, vol. 8, no. 3, p. 57410, 2013.
[47] K. Okamura and S. Yamada, "Calibrating trust in autonomous systems in a dynamic environment," in *Proc. 42nd Annu. Meet. Cogn. Sci. Soc.*, 2020, pp. 1–6.
[48] J. Y. C. Chen and M. J. Barnes, "Human–agent teaming for multirobot control: A review of human factors issues," *IEEE Trans. Human-Machine Syst.*, vol. 44, no. 1, pp. 13–29, Feb. 2014.
[49] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behav. Res. Methods, Instrum., Comput.*, vol. 31, no. 1, pp. 137–149, Mar. 1999.
[50] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in *Proc. 32nd Conf. Neural Inf. Process. Syst.*, 2018, pp. 7549–7561.
[51] S. M. Merritt, K. Huber, J. LaChapell-Unnerstall, and D. Lee, "Continuous calibration of trust in automated systems," Air Force Res. Lab., Wright-Patterson Air Force Base, OH, USA, Tech. Rep. AFRL-RH-WP-TR-2014-0026, 2014.
[52] K. Okamura and S. Yamada, "Calibrating trust in human-drone cooperative navigation," in *Proc. 29th IEEE Int. Conf. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2020, pp. 1274–1279.
[53] *SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, SAE International, Warrendale, PA, USA, 2018.
[54] L. Chen, L. Sun, T. Yang, L. Fan, K. Huang, and Z. Xuanyuan, "RGB-T SLAM: A flexible SLAM framework by combining appearance and thermal information," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2017, pp. 5682–5687.
[55] M. R. Endsley, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics*, vol. 42, no. 3, pp. 462–492, Mar. 1999.
[56] D. B. Kaber and M. R. Endsley, "The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task," *Theor. Issues Ergonom. Sci.*, vol. 5, no. 2, pp. 113–153, Mar. 2004.
[57] J. van Diggelen, J. S. Barnhoorn, M. M. M. Peeters, W. van Staal, M. L. Stolk, B. van der Vecht, J. van der Waa, and J. M. Schraagen, "Pluggable social artificial intelligence for enabling human-agent teaming," in *Proc. NATO HFM Symp. Hum. Auton. Teaming*, 2018, pp. 1–26.
[58] B. van der Vecht, J. van Diggelen, M. Peeters, J. Barnhoorn, and J. per van der Waa, "SAIL: A social artificial intelligence layer for human-machine teaming," in *Proc. Int. Conf. Pract. Appl. Agents Multi-Agent Syst.*, 2018, pp. 262–274.
[59] M. L. Cummings and S. Bruni, "Collaborative human–automation decision making," in *Handbook of Automation*. Cham, Switzerland: Springer, 2009, pp. 437–447.
[60] N. Ezer, S. Bruni, Y. Cai, S. J. Hepenstal, C. A. Miller, and D. D. Schmorrow, "Trust engineering for human-AI teams," in *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 2019, pp. 322–326.
[61] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Hum.–Comput. Interact.*, vol. 36, no. 6, pp. 495–504, Apr. 2020.
[62] M. Kim, E. Seong, Y. Jwa, J. Lee, and S. Kim, "A cascaded multimodal natural user interface to reduce driver distraction," *IEEE Access*, vol. 8, pp. 112969–112984, 2020.

KAZUO OKAMURA received the B.E. and M.E. degrees in information science from Kyoto University, and the Ph.D. degree in informatics from the Graduate University for Advanced Studies (SOKENDAI), in 2020. In 1986, he joined Panasonic and has been involved in research projects and software developments for audio-visual equipment and in-vehicle information systems. He is currently an Executive Engineer at Panasonic Corporation. His research interests include human-agent interaction, artificial intelligence, and human factors in product design.

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