

Adaptive Trust Calibration in Human-AI Cooperation

by

Kazuo Okamura

Dissertation

submitted to the Department of Informatics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy



The Graduate University for Advanced Studies, SOKENDAI

September 2020

Committee

Professor Seiji Yamada (Chair)	National Institute of Informatics and the Graduate University for Advance Studies
Associate Professor Ryutaro Ichise	National Institute of Informatics and the Graduate University for Advance Studies
Associate Professor Kenro Aihara	National Institute of Informatics and the Graduate University for Advance Studies
Associate Professor Tetsunari Inamura	National Institute of Informatics and the Graduate University for Advance Studies
Associate Professor Akihiro Maehigashi	Shizuoka University

Acknowledgments

I would like to express my deepest appreciation to my advisor Professor Seiji Yamada for his guidance, patience, and profound vision regarding the research. It has been a great honor to be his Ph.D. student. Without his continuous encouragement and contagious optimism, I could not have reached this far.

I would like to thank my dissertation committee, Associate Professor Ryutaro Ichise, Associate Professor Kenro Aihara, Associate Professor Tetsunari Inamura, and Associate Professor Akihiro Maehigashi, for their constructive suggestions and invaluable comments, helping me to improve the research.

I would thank Dr. Song Sichao, Ryo Nakahashi, Takato Okudo, Takahiro Tsumura, Yuma Nishi, Nungduk Yun, and the other members of Yamada Laboratory. I greatly appreciate their friendship as well as the good advice and comments on my research.

Sincere thanks go to Yoshiyuki Miyabe, Senior Managing Director of Panasonic, who encouraged me to pursue a doctoral degree.

I am also grateful to Dr. Kuniaki Tatsumi, Dr. Natsuko Sakai, and Dr. Koji Watari from the National Institute of Advanced Industrial Science and Technology for their continuous support. Finally, very special thanks are dedicated to my family for all of the support they showed me during this research, especially when I was living in Tokyo away from home. I would like to thank my wife, Masako, who has made countless sacrifices to help me get through this challenging time in the most positive way. This dissertation would not have been possible without her warm love, continued patience, and endless support.

Abstract

Recent advances in AI technologies are dramatically changing the world and impacting our daily life. The application areas are rapidly expanding, such as autonomous cars, industrial robots, medical services, and various web services. Human users essentially need to cooperate with AI systems to complete tasks 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. The presence and absence of trust definitely impact human behavior and the outcome of cooperation. For optimal performance and safety of human-AI cooperation, the human users must appropriately adjust their level of trust to the actual reliability of AI systems. This process is called “trust calibration”. Users often fail to calibrate their trust properly and end up in a status called “over-trust” or “under-trust” in dynamically changing environments in which an AI’s reliability may fluctuate. Poorly calibrated trust can be a major cause of serious issues with safety and efficiency.

A large number of existing studies on trust calibration emphasize the importance of system transparency to maintain appropriate trust. They claim that appropriate trust could be developed if an AI system provides enough information for a human user to obtain a good understanding of the system. Their primary goal is to avoid over-trust or under-trust, not to deal with improper trust calibration.

Trust is notoriously hard to measure as it is a psychological construct. Self-reported scales of trust that are widely used in most trust literature are too intrusive to use during task executions. Extensive studies have been conducted to examine the factors influencing trust. Although their findings revealed the diversified latent structures of human trust, they suggest that it would be difficult to control human trust intentionally

just by manipulating these factors. Thus, both measuring and influencing trust are challenging issues.

This dissertation focuses on the problem of over-trust and under-trust in human-AI cooperation by exploring two research questions: (1) Can we detect if a user is over-trusting or under-trusting an AI system? (2) Can we mitigate a user's over-trust or under-trust?

We approach the research challenges with a behavior-based trust measurement to capture the status of calibration. Human-AI cooperation is defined as a series of actions taken by a human user and an AI system working on repeated selection problems to decide on either AI execution or manual execution for better performance. A method of adaptive trust calibration is proposed, including a formal framework for detecting improper trust calibration; cognitive cues called "trust calibration cues"; and a technical architecture of human-AI cooperation with a concept called trust calibration AI.

Three empirical studies were done to evaluate the proposed method. We designed two experimental tasks for human-AI cooperation: a pothole inspection task and a continuous cooperative navigation task. We conducted three online experiments using a simulated drone environment. The results of the first empirical study demonstrate that our proposed method has significant effects on changing human behavior in the case of over-trust. The second empirical study shows that the proposed method also works well under dynamic trust changes of ABA and BAB, where A and B mean over-trust and under-trust. The third empirical study indicates that the proposed method is effective in a continuous real-time task involving navigating a semi-autonomous drone. This result can open the possibility of applying the proposed method to practical real-time applications such as autonomous driving. We also discuss a possible extension to the framework with expected utility functions to incorporate trust factors other than performance.

The results of the empirical evaluations indicate that the proposed method could detect and mitigate the status of improper trust calibration; therefore, we conclude that our proposed method provides a reasonable basis for answering the two research questions. As the proposed method is based on a simple and task-independent framework, it could be applied to many application situations. Despite several limitations, this dissertation contributes to providing a basic framework for managing trust calibration, leading to better interaction designs for human-AI cooperation.

Contents

List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Background	1
1.2 Research Question and Approach	3
1.3 Outline	4
2 Related Work	7
2.1 Conceptualizing Trust	7
2.2 Factors Influencing Trust	8
2.3 Modeling Trust	11
2.4 Measuring Trust	14
2.5 Trust Calibration	17
2.6 Trust Research in Human-Robot Interaction	20
3 Adaptive Trust Calibration	23
3.1 Introduction	23
3.2 Framework for Detecting Over-Trust and Under-Trust	24
3.3 Trust Calibration Cue	25
3.4 Method of Adaptive Trust Calibration	26
4 Empirical Studies	29
4.1 Overview	29

4.2	Evaluation with an Over-trust Scenario	32
4.2.1	Introduction	32
4.2.2	Method	33
4.2.3	Results	40
4.2.4	Discussion	44
4.2.5	Conclusion	48
4.3	Evaluation under Bi-directional Trust Changes	48
4.3.1	Introduction	49
4.3.2	Method	49
4.3.3	Results	56
4.3.4	Discussion	61
4.3.5	Conclusion	64
4.4	Evaluation with Continuous Cooperative Tasks	64
4.4.1	Introduction	64
4.4.2	Method	65
4.4.3	Results	72
4.4.4	Discussion	74
4.4.5	Conclusion	76
5	Conclusion	77
5.1	General Discussion	77
5.1.1	Applicability of Proposed Method	77
5.1.2	Role of Trust Calibration AI	79
5.1.3	Extension to Framework	81
5.2	Conclusion	84
5.3	Future Work	86
	Bibliography	87
	List of Publication	103

List of Figures

2.1	Conceptual model proposed by Lee and See [1]	12
2.2	Trust model defined in Hoff and Bashir [2]	14
2.3	A conceptual organization of trust and human-automation interaction used by Drnec et al. [3]	15
2.4	Trust behavior model defined by Bindewald et al. [4]	16
2.5	Representation of trust calibration (redrawn from Lee and See [1]) . . .	17
2.6	Driver's information module (left). Indicator of car's ability to drive autonomously (right) (Adapted from Helldin et al. [5])	19
2.7	Visualization of autonomous car's interpretation of current situation: world in miniature (left), chauffeur avatar (middle), and display of car's indicator (right) as baseline (Adapted from Haeuslschmid et al. [6]) . .	19
3.1	Human-AI cooperation	27
3.2	Human, Task-AI, and TCAI	28
4.1	Four types of TCCs.	32
4.2	Online drone simulator.	34
4.3	Popup message asking the participants for choice.	35
4.4	Reliability Indicator at the bottom left area of the screen.	36
4.5	Popup windows of the pothole inspections.	37
4.6	Relationship between the three parameters under changing weather conditions	39
4.7	Manual rates over time	42
4.8	Manual rates for each TCC during the bad weather period	43

4.9	Verbal TCC	50
4.10	The instruction screen	53
4.11	P_A and P_H in the ABA scenario	54
4.12	TCC rates of TCC-ABA group	56
4.13	TCC rates of TCC-BAB group	57
4.14	Manual rates	60
4.15	Sensitivity d'	61
4.16	Online semi-autonomous drone simulator	66
4.17	The first part of the course	67
4.18	Cross-track error	68
4.19	Verbal TCC	68
4.20	TCC rates	72
4.21	Manual-pilot rates	73
4.22	Cross-track errors	74
5.1	Three categories of factors influencing trust decisions	81

List of Tables

4.1	Three empirical studies.	30
4.2	Comparisons of cooperative tasks used in empirical studies.	31
4.3	Means of TCC rates at each CKP.	41
4.4	Means of the other dependent measures	41
4.5	3-CKP mean values of P_H	44
4.6	3-CKP means of TCC rates in each condition	58
4.7	Means of the manual rates and the sensitivity d'	58
5.1	Typical applications and their compliance with requirements	80

1

Introduction

This chapter introduces the topic of this dissertation and provides an overview of it. Section 1.1 introduces the research background and section 1.2 presents the research questions and describes the approach. Section 1.3 gives the structure of the dissertation.

1.1 Background

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 [7, 8]. The presence and absence of trust definitely impacts human behavior and the outcome of cooperation [9, 10, 11]. 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 [1].

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” [8]. 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 in which the user overestimates 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. An official report [12] on a fatal crash of an autonomous vehicle in California concluded that one of the probable causes of the accident was the driver’s over-trust in the vehicle’s driving automation system. Similar car accidents caused by over-trust have also been reported [13, 14]. Under-trust is poorly calibrated trust in which the user underestimates the AI’s reliability; it can result in an agent not being used, excessive user workload, or deterioration in the total system performance [15].

In keeping appropriate trust, it is necessary to be able to 1) measure trust and 2) influence trust if necessary. However, these two elements are still challenging issues.

Measuring trust is not easy, as trust is a latent construct. Most of the research on trust has used self-reported trust scales; however, they are so intrusive that it is 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 [16]. 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.

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, 18, 19]. Hoff and Bashir [2] reported 29 factors that are influential in the development of human trust. Schaefer et al. [9] 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.

Most of the existing research on trust calibration such as [20, 21, 5, 22, 6] emphasizes the importance of system transparency to maintain appropriate trust. The authors in these examples claim that appropriate trust could be developed if an AI system provides enough information for a human user to obtain a good understanding of the system. The categories of information necessary to provide better system transparency are essentially inline with the factors influencing human trust. The primary goal is to avoid trust miscalibration. Although recent works such as [15, 23] proposed trust calibration models for human-robot teams, not many studies have focused on how to detect improper trust calibration nor how to mitigate it.

1.2 Research Question and Approach

With the challenges described above in mind, this dissertation focuses on the problem of over-trust or under-trust by exploring two research questions:

- **RQ1 : Can we detect if a user is over-trusting or under-trusting an AI system?**

To address this question, first, it would be necessary to define trust calibration in human-AI cooperation formally. Such a definition should be able to describe the status of over-trust and under-trust. Second, a method of capturing changes in calibration status would be required. It would be desirable that such a method be realized without directly measuring trust.

- **RQ2 : Can we mitigate a user's over-trust or under-trust?**

Once human users fall into the categories of over-trust or under-trust, it might not be easy for them to escape from the status of miscalibration. Calibration can only occur in response to new evidence that may change the user's awareness, while no new evidence can be learned without changing the current behavior first [24]. A method for escaping from this vicious cycle should be proposed. Such a method should be evaluated to determine whether it can help users calibrate their trust appropriately to improve performance in human-AI cooperation.

In this dissertation, we define human-AI cooperation as a set of repeated actions by human users and AI systems working together to perform a common task interchangeably in order to achieve a better performance outcome. Examples of applications include autonomous driving, medical diagnostic assistance, and a baggage screening system. Among the various factors influencing trust, this dissertation mainly focuses on performance-related factors. One of the fundamental goals of human-AI cooperation is to achieve higher performance than what humans and AI can achieve independently. Consideration of other influencing factors is also discussed later.

To address the research questions, we defined a framework for describing the trust calibration status in human-AI cooperation, combined with a behavior-based measurement for detecting improper trust calibration. We also examined several types of cognitive cues for notifying users of the status of improper trust calibration. Empirical evaluations of the proposed method were conducted as online experiments using a semi-autonomous drone simulator, with two types of cooperative tasks and under different environmental conditions.

1.3 Outline

This dissertation is organized as follows:

Chapter 2: Related Work This chapter reviews existing trust research literature on automation and computer-human interaction, mainly focusing on the calibration of trust in AI or autonomous systems.

Chapter 3: Adaptive Trust Calibration. This chapter proposes a method of adaptive trust calibration. The proposed method consists of a formal framework for defining the status of improper trust calibration and cognitive cues called “trust calibration cues” for notifying users of the miscalibration status.

Chapter 4: Empirical Studies. In this chapter, three empirical studies done to evaluate the proposed method are presented and discussed. First, an overview of the empirical studies is explained. The first study is done to verify how the proposed method works in a simple over-trust scenario using four types of trust calibration cues.

The second study evaluates the proposed method in bidirectional trust changes of ABA/BAB scenarios of under-trust (A) and over-trust (B). The target task of human-AI cooperation in the first two studies is pothole inspection in which human users check if there are holes or cracks in road images from a drone simulator. The users may use automatic inspection or do the inspection by themselves. To examine how effective the proposed method would be in real-time applications such as autonomous driving, a third empirical study is done using a cooperative navigation task with a semi-autonomous drone. In contrast to the fact that the first two empirical studies use a set of discrete and independent tasks involving pothole inspection, the third study evaluates the proposed method with a series of continuous and interdependent tasks involving drone navigation.

Chapter 5: Conclusion. This chapter discusses the studies presented in the previous chapters, highlights the contributions of this dissertation, and recommends areas for future research.

2

Related Work

This chapter gives the related work on trust research to provide the background of the dissertation.

2.1 Conceptualizing Trust

Trust and its role in mediating the relationship between humans and computers has been widely recognized. Many studies have been done in the field of aviation, automation, and human computer interaction. There are many types of definitions of trust due to the complexity and the multi-faceted nature of trust.

One of the accepted definitions can be found in organizational theory, which views trust as an intention: “a willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party” [25]. In this definition, trust concerns an expectancy or an attitude regarding the likelihood of favorable responses from a trustee. A similar definition can be found in [26].

Just as in interpersonal relationships, trust plays a vital role in determining humans' willingness to rely on autonomous systems, even in situations of risk or uncertainty. Madsen and Gregor [27] define human-computer trust as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid." This definition covers both a user's confidence in a system and their willingness to act on the system's decisions and advice. Lee and See [1], in their influential review of research on trust and reliance, introduce a basic definition of trust: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." This definition is often used in empirical studies of trust in automation, although a widely accepted definition of trust is still lacking [28].

Two types of trust definitions are particularly relevant to this dissertation. First, we adhere to one definition of trust [19], "expectation related to subjective probability an individual assigns to the occurrence of some set of future events." One of the fundamental goals of human-AI cooperation is achieving better performance as a result of successful events. Second, we are also interested in the alternative definition of trust described in Adams et al. [29] to view trust as observable choice behavior. This dissertation focuses on trust in human-AI cooperation in which human users need to choose whether they should rely on an AI counterpart or not.

2.2 Factors Influencing Trust

Factors influencing trust in autonomous systems are commonly divided into three categories [29, 30]: human characteristics, characteristics of the autonomous system, and environmental factors. Although some studies indicate that the characteristics of the autonomous systems have the largest impact on trust [31, 9], human-related factors and environmental factors are also essential to understanding the complexity of trust formation in real-world scenarios.

Human-related Factors

- **Culture**

The way that a technology is accepted in culture, either as positively or negatively,

significantly affects how people use the technology. Related research can be found in [32, 33]

- **Age**

Several studies such as [34, 35] revealed that there are age-based differences in human trust. They also suggested that people of different ages may differ in their trust assessment strategies and that the effects of age on trust depend on context.

- **Personality**

Research has shown that people with a high propensity to trust are more likely to trust autonomous systems than those with a lower propensity, although the effects of personality on trust may depend on the functions and tasks of the autonomous systems. Related research can be found in [36, 37].

- **Understanding of system, and expert knowledge**

Human users may have different abilities in terms of understanding how autonomous systems work as based on prior knowledge and expectations. Balfe et al. [38] observed real railway operators using an automated train-route setting system. They reported that understanding the automated system was a stronger dimension in trust development than the system's reliability or competence. The work of Sanchez et al. [39] demonstrated that experts with prior knowledge on a specific domain often show an unwillingness to rely on automation when a system is faulty. They argued that any error would have a stronger negative impact on perceived reliability if expectations toward automation are higher.

- **Self-confidence**

Early works [40, 41] suggested a simple relationship between trust and self-confidence in which automation would be used when trust was higher than self-confidence, and manual control would be used when self-confidence exceeded trust. Lewandowsky et al. [10] also suggested that the difference between trust and self-confidence was a strong predictor of a human's reliance on automation. They argue that self-confidence has a stronger effect on human-automation collaboration than interpersonal collaboration since a human operator would be solely responsible for the task results in the former case while human operators could share responsibility in the latter case.

System-related Factors

- **System reliability**

A great deal of empirical research such as [42, 43, 44] shows that reliability is closely related to trust. A higher reliability could lead to higher trust [45]. Declining system reliability over time could lead to decreasing trust [46]. The reliability of an automated system is a strong predictor of trust in the system.

- **System failure** System failure, which is a form of system reliability, specifically refers to discrete erroneous events within a system. Previous studies demonstrated that system failure has a negative impact on trust in automation [29]. Once system failure occurs, trust often decreases drastically. After a failure, the levels of trust recover slowly and often do not return to previous levels [47]. Yu et al. [48] revealed that system failures have a stronger effect on trust than system successes. de Vires et al. [41] showed that large failures have a more negative effect on trust than small failures. Muir et al. [11] demonstrated that small errors that vary in magnitude reduced trust more than large constant failures.

- **System transparency** Systems transparency refers to accurate feedback concerning the reliability and situational factors that can affect the reliability of a system [2]. Systems that can explain their reasoning are more likely to be trusted, as this facility can make system functioning more easily understood [49].

- **System predictability** Predictability will also impact trust. An autonomous system that can be predicted to have reliable and consistent performance is more likely to be trusted [29]. Increasing user familiarity with a system decreases the rate of trust change during system failure [50].

- **Levels of autonomy** The levels of autonomy, which describe task allocation between human and automated systems, may complicate the development of trust in a system since higher levels of autonomy are generally harder for humans to understand. Recent advances in artificial intelligence are accelerating this issue.

Environment-related Factors

- **Risk**

Risk is known to impact trust in automation [51, 16]. Trust permits a trustor to act in a manner that puts them at considerable risk, believing that the actions of their counterpart will mitigate that risk [52]. Reliance on automation is moderated by the level of risk in an interaction [28].

- **Workload and task**

Biros et al. [53] found that users of an automated decision support system rely on automation more in the case of high task workloads, regardless of their level of trust. Positive correlations between trust and reliance are impacted by workload. One of the benefits of introducing automation other than better performance is that it can allow the user to perform another task. This dual-task aspect has been well examined in autonomous driving research such as [54, 55] trying to balance safety and convenience in performing non-driving-related tasks.

This dissertation mainly examines the trust factors related to the performance of both human users and AI systems, such as self-confidence, system reliability, and system transparency.

2.3 Modeling Trust

Many theoretical and quantitative models of trust in automation or autonomous systems have been proposed to integrate the factors influencing trust in autonomous systems.

Lee and See [1] studied how the characteristics of autonomous systems and human cognitive processes affect the appropriateness of trust, which guides reliance when complexity and unanticipated situation make a complete understanding of the autonomous system difficult. They discussed trust development factors in the dimensions used in Lee and Moray[47]: performance, process, and purpose. Performance refers to the operation of autonomous systems such as reliability, predictability, and ability. Process describes how appropriate the algorithms are for achieving the goals of the system. Purpose describes the designer's intents of the system. A conceptual model

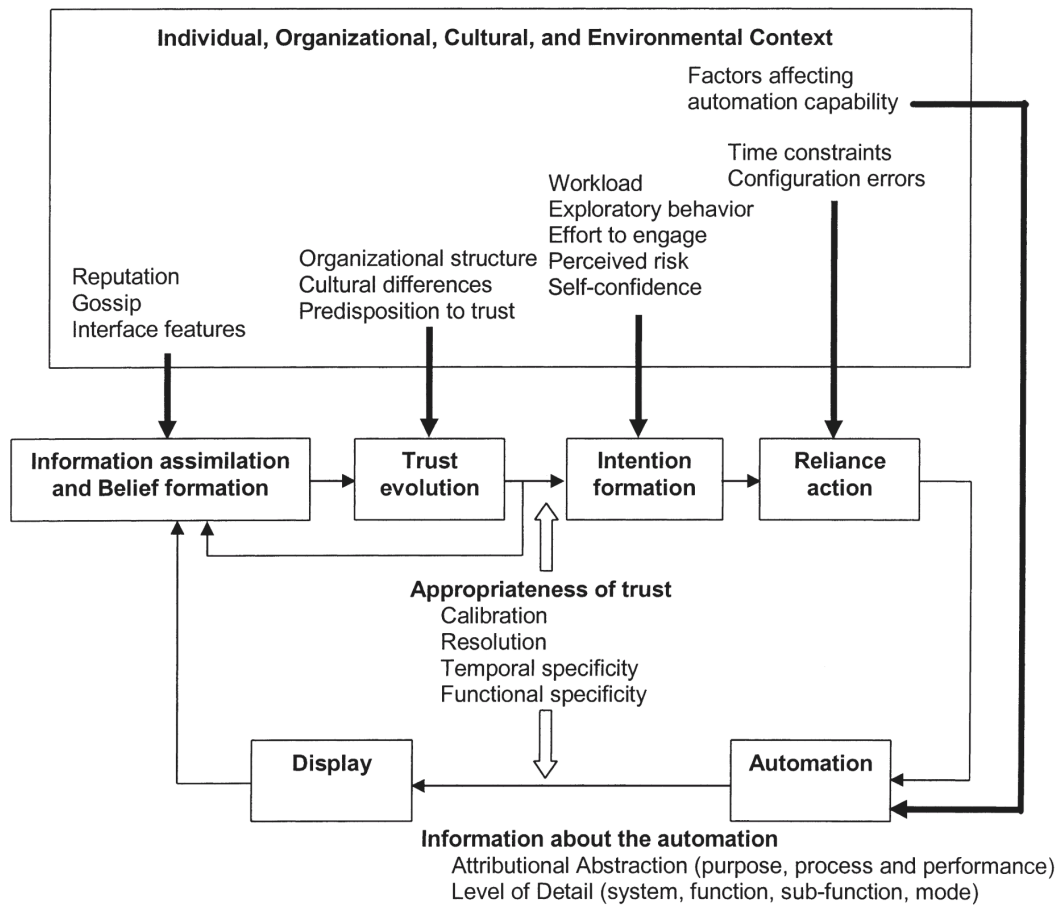


Figure 2.1: Conceptual model proposed by Lee and See [1]

(Figure 2.1) was proposed to integrate the various factors in the dynamic process of trust development and reliance action. They claimed that the appropriateness of trust depends on the information that a human can get by observing a system. This type of information is known as system transparency, described in the following section in this chapter.

Hoff and Bashir [2] conducted a systematic review of 127 pieces of empirical research on trust in automation. The types of automation systems in the reviewed research include combat identification aid, a decision support system, a luggage screening system, collision warning, and a route-planning system. Figure 2.2 shows the

results of their review as a three-layered trust model for categorizing the factors that influence trust in automation: dispositional trust, situational trust, and learned trust. Dispositional trust represents an individual's overall tendency to trust. Situational trust depends on the specific context of an interaction. The environment has a strong influence on situational trust, and human mental states can also alter situational trust. The final layer, learned trust, which is knowledge on a system drawn from past experiences or a current interaction, is further divided into two types: initial learned trust, that is, trust prior to interacting with a system, and dynamic learned trust, that is, trust formed during an interaction. This model covers the dynamic nature of the trust process both prior to an interaction and during an interaction. They also presented in their model that both initial and dynamic learned trust, which greatly influence reliance on automation, are not the only contributing factors. Additional factors on the human side, such as efforts made to use a system, situational awareness, and physical state, also impact the reliance decision. Figure 2.2 is a diagram of their model.

Hancock et al. [31] examined factors that affect trust in human-robot interaction (HRI) by applying meta-analysis methods to existing empirical studies. They classified factors of trust development into three broad categories according to the experimental manipulation: robot-related factors (including performance-based and attribute-based factors), human-related factors (including ability-based and human characteristic factors), and environment-related factors affecting trust (including team collaboration and task-based factors). A meta-analysis of 27 studies with these factors provided 69 correlational and 47 experimental effect sizes. They found that the robot performance and attributes were the factors most associated with trust in human-robot interaction.

Drnec, Metcalfe, and their colleagues [3, 56] insisted on the benefit of re-framing the trust problem space into a behaviorally defined one with reliance and compliance. This would remove the need for inferencing latent constructs such as trust. Figure 2.3 is their conceptual diagram adapted from [57], which highlights three major elements in their model: trust in automation (TiA), interaction decision, and trust outcomes or behavior.

Bindewald et al. [4] developed a model (see Figure 2.4) to describe the relationship between trust and reliance by showing causal links.

Several other studies [41, 58, 50] also investigated the relationships among trust, self-confidence, and reliance behaviors. Our proposal in this dissertation is inspired by

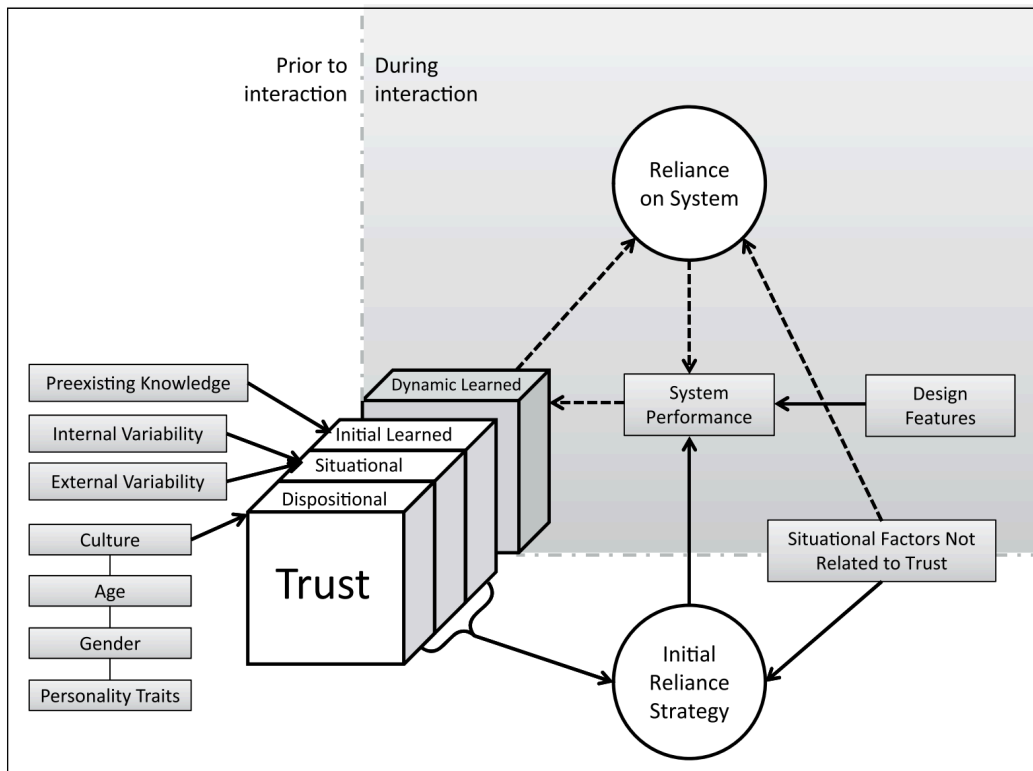


Figure 2.2: Trust model defined in Hoff and Bashir [2]

these findings on the structures behind human trust behaviors.

2.4 Measuring Trust

Measuring trust is difficult in general, as trust is a latent construct. Most of the previous trust literature used self-report measures such as subjective rating scales with questionnaires. Jian et al. [59] developed a scale for measuring trust in automation, which is commonly used in trust research literature. They conducted three empirical studies to define a 7-point scale for 12 statements, which almost correspond to three dimensions: benevolence (purpose), integrity (process), and ability (performance). Madsen and Gregor [27] developed a Human-Computer Trust scale that consists of five constructs: reliability,

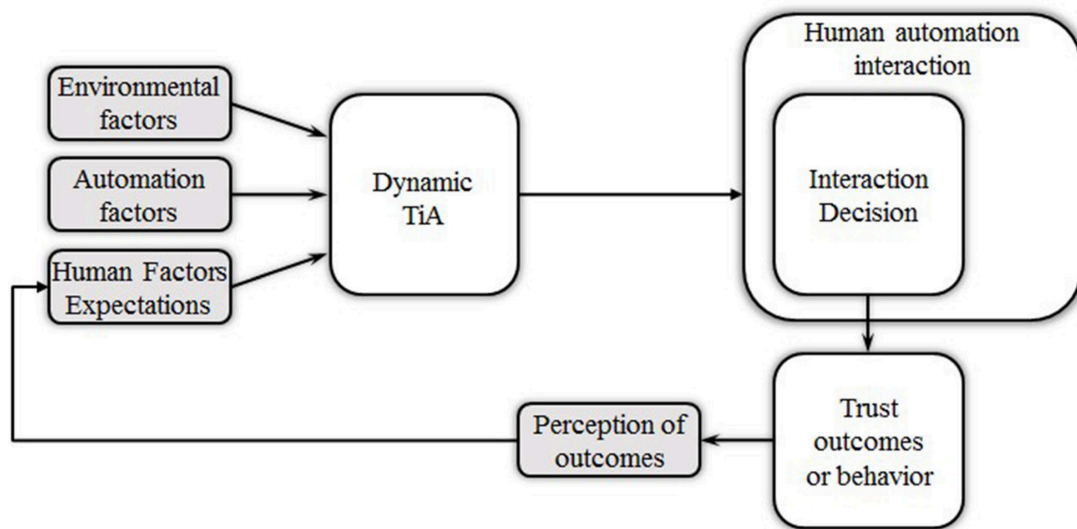


Figure 2.3: A conceptual organization of trust and human-automation interaction used by Drnec et al. [3]

technical competence, understandability, faith, and personal attachment. A total of 25 statements is used to evaluate trust in terms of “cognitive based” and “affective based.” The SHAPE Automation Trust Index [60] was developed for practical use to measure trust in air traffic control systems. This scale consists of seven dimensions: reliability, accuracy, understanding, faith, liking, familiarity, and robustness. Yagoda and Gillan [61] developed the Human-Robot Interaction Trust Scale with five dimensions: team configuration, team process, context, task, and system. The disadvantages of the self-report measures are too intrusive and cannot be used during task execution. Questionnaires taken at the end of experiments sometimes do not correctly reflect real-time trust [16].

Many trust models, including the ones explained in the previous section, define the relationship between trust and behaviors of reliance or compliance. Behavioral trust measures have a theoretical basis in such models and can be consistently used during task execution; therefore, they can be useful in real-world applications. Behavioral measures can also serve as a basis for modeling and prediction [3]. Behavior used as trust measures includes selecting manual or automatic tasks [41, 62], choosing an automation level [16], reaction time [63], and accepting advice (for reliance or

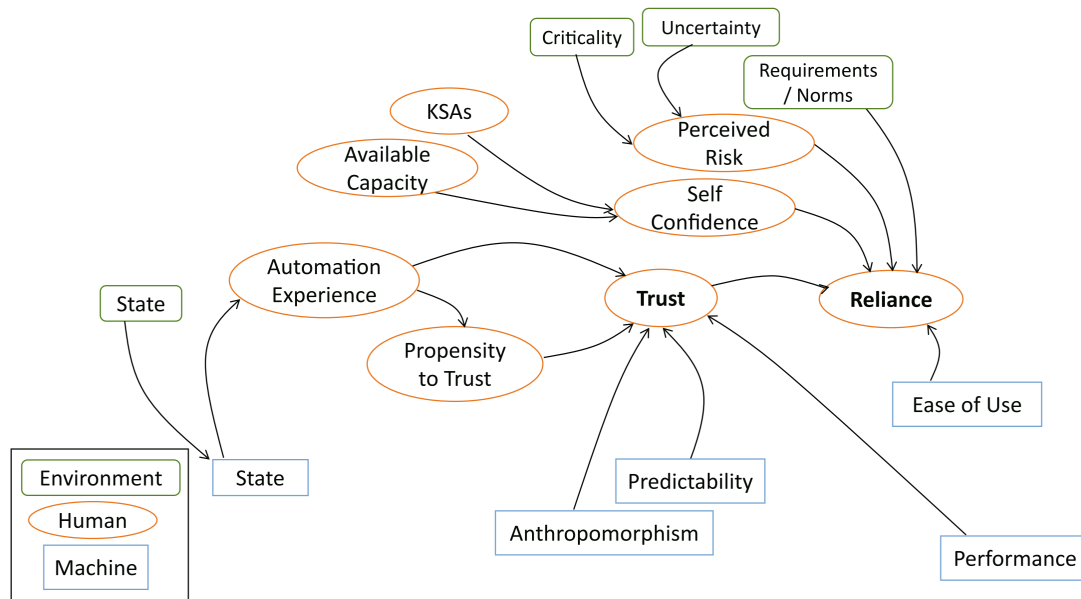


Figure 2.4: Trust behavior model defined by Bindewald et al. [4]

compliance)[64, 65, 4, 66].

Hergeth et al. [67] demonstrated that there was a significant negative correlation between participants' trust in automation and the monitoring frequency calculated from gaze behaviors. Other types of physiological and neural measures include facial and voice tracking [68], heart rate, and EEG. Although these measures can be used for the dynamic tracking of trust, they usually require special hardware. Further research would be necessary to clarify the correlation between trust and these metrics.

Instead of direct measurement, inferring trust is another approach. Lee and Moray [47] made an early attempt. They proposed a temporal model for relating trust in automation to task performance factors, using an auto-regressive and moving average value regression approach. Xu and Dudek [69] proposed a trust inference model called "OPTIMO," which is a dynamic Bayesian network over human moment-to-moment trust states, based on robot task performance and operator reactions over time. Chen et al. [70] demonstrated a POMDP-based trust model to infer human trust through interaction with a robot. Other related studies can be found in [71, 72].

Considering the challenge of measuring trust, given that it is a complex psychologi-

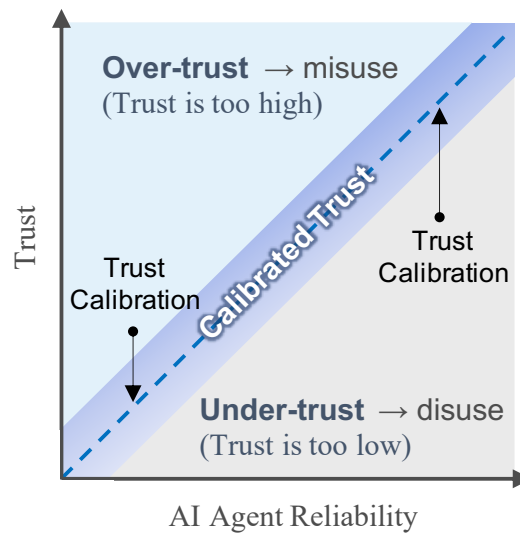


Figure 2.5: Representation of trust calibration (redrawn from Lee and See [1])

cal construct, our proposal in this dissertation takes a behavior-based approach to estimate the status of trust calibration.

2.5 Trust Calibration

Human decisions to rely on AI agents directly affect the total performance of human-AI cooperation. Human users who have greater trust in an agent tend to rely more on it than users with lower trust [47, 37, 73]. Appropriate reliance decisions depend on how well users adjust their level of trust to an agent's actual reliability. This process is called *trust calibration* [8]. Maintaining appropriate trust is critical in avoiding misuse and disuse [74, 75]. Over-trust is poorly calibrated trust in which the user overestimates the reliability of the agent; it can result in misuse of an agent with the expectation that the agent can perform outside of its designed capability. Under-trust is poorly calibrated trust in which the user underestimates the agent's capability; it can result in an agent not being used, excessive user workload, or deterioration in the total system performance. Poor trust calibration sometimes causes serious safety issues [76, 77, 14, 13]. Figure 2.5 illustrates the relationship between trust and AI agent reliability. The diagonal line represents the calibrated trust where the level

of trust matches the agent reliability. Above this line means over-trust, and below implies under-trust. Lee and See [1] suggested two additional concepts to discuss trust calibration: resolution and specificity. Resolution refers to how accurately a judgment of trust discriminates different levels of reliability. Specificity means the extent to which trust is associated with a particular component of an agent. While an agent's reliability changes for various reasons in an environment, users often fail to calibrate their trust in the agent and end up over-trusting or under-trusting it. Merritt et al. [78] also proposed three metrics for evaluating trust calibration. They distinguished trust and perceived reliability in their study. Perceptual accuracy refers to the degree to which the user perceives an agent's reliability accurately. Perceptual sensitivity and trust sensitivity reflect how users adjust perceived reliability and trust as the agent reliability changes over time.

One of the keys to maintaining appropriate trust calibration is the concept of system transparency. Humans require a user interface that captures the state of an entire system in order to interact appropriately with an agent. System transparency has been defined as “the quality of an interface pertaining to its ability to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process [79].” This definition is essentially in accordance with the factors that influence human trust [1]: purpose, process, and performance. Although system transparency reduces uncertainty in AI agents, it can also create more uncertainty when done poorly [80].

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. [20] 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 [5, 22, 6] have indicated that providing good transparency by constantly presenting system information helps maintain the appropriate trust in vehicles.

Helldin et al. [5] 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 (see Figure 2.6) 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



Figure 2.6: Driver's information module (left). Indicator of car's ability to drive autonomously (right) (Adapted from Helldin et al. [5])

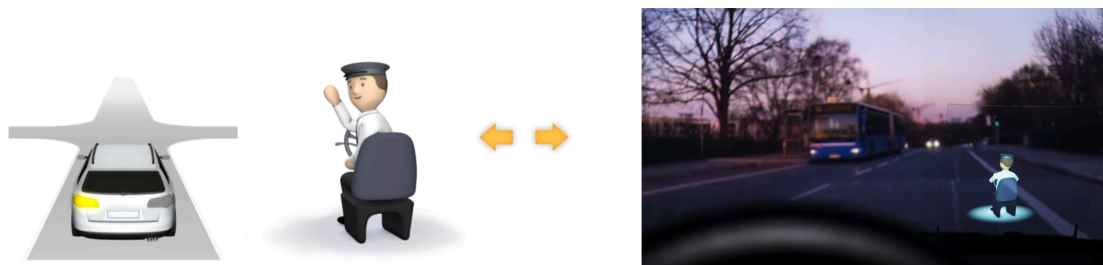


Figure 2.7: Visualization of autonomous car's interpretation of current situation: world in miniature (left), chauffeur avatar (middle), and display of car's indicator (right) as baseline (Adapted from Haeuslschmid et al. [6])

information also took control of the car faster when needed and were able to perform tasks other than driving without risking safety. Haeuslschmid et al. [6] compared three different visualizations of an autonomous car's interpretation of the current situation (see Figure 2.7): a world in miniature, a chauffeur avatar, and a display of the car's indicator as the baseline. They found that the world in miniature visualization increased trust the most. Seppelt [81, 82] examined drivers' trust in adaptive cruise control with or without the display of continuous information that provides information on system behavior in real-time. They found that that continuous feedback on automation behavior viably promotes calibrated trust and reliance. Mercado et al. [83] conducted experiments with operators of unmanned vehicles to evaluate the impact of different levels of transparency, including information such as planned paths and goals on a map

and verbal explanations of the rationale behind agent reasoning. Their results revealed the benefits of transparency in terms of effective performance without additional workload costs. Using a high-fidelity simulated environment of an automated aviation aid system, Lyons et al. [84] demonstrated that the use of logic-based explanations for recommendations was found to promote a pilot's trust in an automated system. Other related studies can be found in [85, 71, 86, 75, 87].

In the research areas of artificial intelligence and also human-robot interaction, there is growing interest in explainability, which is one of the components of system transparency as described above. Complex AI systems need to explain how they work so that users can develop appropriate trust [49]. Related studies on explainability and trust can be found in [88, 89, 90, 91, 92, 93].

Most of the studies investigated how to maintain appropriate trust calibration by continuously presenting system information to prevent over-trust or under-trust. Only a few works such as [94, 15] in the literature have focused on detecting poor trust calibration or how to recover from over-trust or under-trust swiftly.

2.6 Trust Research in Human-Robot Interaction

Many studies have been done on human-robot interaction regarding human trust in robots [31].

Robinette et al. [77] reported that low reliability does not always mean low trust if other factors such as robot appearance and behaviors could moderate the effect of reliability in the case of an emergency. M. Siegel et al. investigated how to build persuasive robots that can change others' beliefs such that such robots are trustworthy [95]. They focused on the genders of a human and a robot in HRI and found that the effect was much stronger between male participants and a female robot in establishing trust.

Sean Ye et al. [96] also investigated the effect that four forms of communication have on trust with and without the presence of robot mistakes during a cooperative task. They found that participants' trust in a robot was better preserved when that robot offered advice. A method of establishing appropriate trust on the basis of a robot's actions from its policy, especially in *critical* states, has been studied [97]. This research asserts that if a robot shows a human what its understanding of the task's

critical states are, the human can make a more informed decision about whether to deploy the policy. Sound formalization for critical states was developed with a Markov decision process, and a demonstration in a highway driving task was shown.

Fast adaptation with meta-reinforcement learning was proposed to establish trust between human users and assistive robots [98]. This work showed that the adaptation of a robot was an important function for gaining a lot of trust from a human user. The results of a simple simulation experiment supported this assertion. D. P. Losey and D. Sadigh proposed a way to exploit human trust in a robot for efficient human-robot coordination [99]. Their concept of “trust” assumes that a human believes that a robot will behave *rationally* toward its objective. In user studies, they showed that trusting human models could lead to communicative robot behavior, which increased their involvement.

Almost all of the trust studies in HRI have been done on robot’s functions for building human trust. In contrast with these, we propose a method for detecting over/under-trust and having humans calibrate their trust by themselves.

3

Adaptive Trust Calibration

In this chapter, we propose a novel method of adaptive trust calibration. In section 3.2, we define a formal framework for detecting over-trust and under-trust. Cognitive cues called “trust calibration cues” (TCCs) are also proposed section 3.3. Section 3.4 describes a method of adaptive trust calibration using the proposed framework and TCCs.

3.1 Introduction

Many types of cooperation are possible between human users and AI in terms of roles and responsibilities. We 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 humans and AI should have the same functionality to execute a common task with different performances depending on the situation. A human user must decide whether a task should be done by AI or humans. The final responsibility for an outcome always belongs to the human user in this type of

cooperation. Examples of this type of cooperation include autonomous driving (driver and AI), medical diagnostic assistance (doctor and AI), and baggage screening systems (airport clerk and AI).

We emphasize two important aspects of trust in human-AI cooperation: *performance* and *human behavior*. 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 [31]. Therefore, we focus on the performance-related factors that influence trust. This focus makes it possible to narrow down the definition of trust as “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 also can be viewed as a human user’s behavior in choosing [29] whether to rely on an AI system or to do each 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.

3.2 Framework for Detecting Over-Trust and Under-Trust

We propose a framework for detecting an inappropriate trust-calibration status with a behavior-based approach. Suppose a user and an AI system are jointly working on a set of tasks. The user should decide whether to rely on the system or do each task manually. In this framework, three 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.”

P_A varies depending on the conditions of the AI system. \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 \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 as follows:

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

$$(\widehat{P}_A > P_H) \wedge (P_H > P_A) \quad (3.1)$$

- **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.

$$(\widehat{P}_A < P_H) \wedge (P_H < P_A) \quad (3.2)$$

Several studies [41, 58, 50] 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 the task manually. Maehigashi et al. [100] found that human users select an automation task or a manual task on the basis of their perceived manual performance. When a user decides to rely on a system, it is reasonable to say that this behavior indicates $\widehat{P}_A > P_H$. If the user decides to do the task manually rather than rely on the system, it indicates that the user estimates $\widehat{P}_A < P_H$. Instead of directly measuring \widehat{P}_A or P_H , the first terms of (3.1) and (3.2) can be estimated by observing the user's reliance behavior; thus, the trust calibration status can be detected if the second terms of P_H and P_A can be estimated.

3.3 Trust Calibration Cue

To effectively notify human users of their improper trust calibration, we explore the idea of giving cognitive cues to users when over-trust or under-trust is detected. Once users fall into the categories of over-trust or under-trust, it might not be easy for them to escape. This could be an example of confirmation bias in the preservation of trust [101]. Calibration can only occur in response to new evidence that may change the users' prevailing awareness, while no new evidence can be learned without changing the current behavior first [29]. To solve this dilemma, which could perpetuate a status of inappropriate trust calibration, a new trigger would be necessary for instances of

over-trust or under-trust. This cue is expected to trigger the user to promptly notice that something has been changed in the environment and to calibrate trust on the basis of the new findings. We call this cognitive cue a “trust calibration cue” (TCC).

Several trust studies have proposed enhancing system transparency by using “cues.” Visser et al. [21] proposed a design guideline for trust cues, which are information elements used to make a trust assessment about autonomous systems. They classified the cues in terms of trust dimensions (intent, performance, process, expressiveness, and origin) and trust processing stages (perception, comprehension, projection, decision, and execution). Cai and Lin [102] examined multi-modal cues for conveying the confidence of a driver assistance system. Unlike our TCC, the goal of these “cues” was to bring system information to users.

The concept of the TCC was inspired by the works done by Komatsu et al. [103], which proposed an intuitive notification methodology called “artificial subtle expressions” (ASE). One of the design requirements is “complementary,” which means that notifications should not interfere with the main communication protocol. A TCC should also be recognized as a notification through a different channel than the one being used as the system-transparency interface of an AI system. A user with improper trust calibration is probably having difficulty in understanding the system information of an AI system.

Specific designs and evaluations of TCCs are described in the next Chapter.

3.4 Method of Adaptive Trust Calibration

With the framework and TCCs described above, we propose a method of adaptive trust calibration.

Figure 3.1 shows a conceptual diagram of human-AI cooperation, which can be viewed as a series of actions taken by a human user working 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 on the basis of P_H and \widehat{P}_A , his/her reliability, and the estimated reliability of the Task-AI. Each decision corresponds to the first inequalities in the

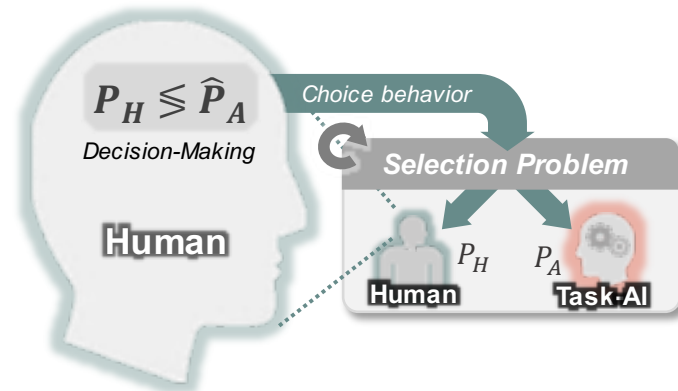


Figure 3.1: Human-AI cooperation

proposed framework, (3.1) and (3.2).

To detect and mitigate improper trust calibration, we introduce a conceptual entity in human-AI cooperation called “trust calibration AI” (TCAI). It is a meta-level entity that manages the whole process of adaptive trust calibration. Figure 3.2 illustrates the relationship among the human user, the Task-AI, and the TCAI.

The TCAI observes human’s choice behaviors which indicate the answers of the selection problems. This observation is to evaluate the first inequalities in the proposed framework. The TCAI also solves the selection problems by estimating P_A and \hat{P}_H with a model-based or statistical approach. These estimations correspond to evaluating the second inequalities in the proposed framework. 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 definitions of the proposed framework, and it gives a TCC to the human user to notify an improper trust calibration status.

The TCAI can solve the selection problems with knowledge of the Task-AI implementation and the human user’s capability for the task execution; however, 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 the human-AI cooperation. The TCAI only suggests to the human user recalibrate trust in the Task-AI by presenting a TCC if the human behavior signifies the improper trust calibration.

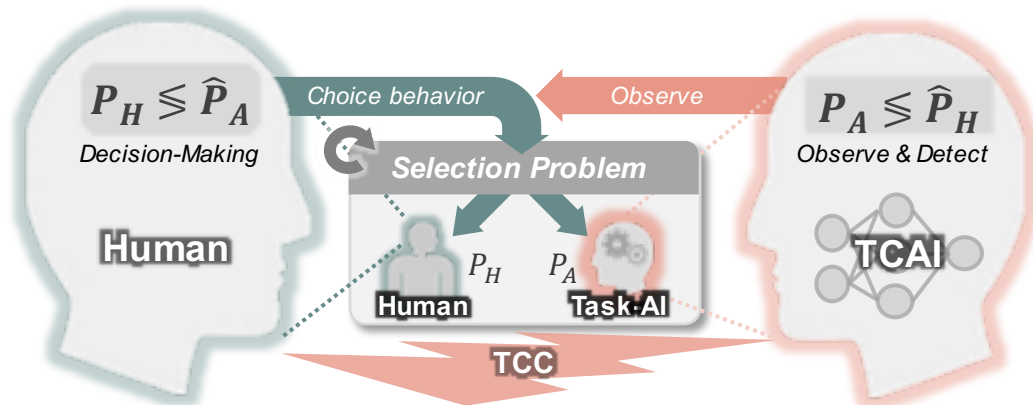


Figure 3.2: Human, Task-AI, and TCAI

The basic algorithm of the method of adaptive trust calibration is described in Algorithm A0. This method aims to adaptively prompt a user to calibrate her/his trust by presenting a trust calibration cue only when our framework 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 : Method of adaptive trust calibration

```

while Cooperative tasks exist do
  Observe a user's choice behavior.
  Evaluate the second inequalities of the framework (3.1) and (3.2).
  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

```

4

Empirical Studies

This chapter presents three empirical studies done to examine how effective the proposed method would be in various trust conditions with two different cooperative tasks.

Section 4.1 describes an overview of the three empirical studies. In section 4.2, the evaluation with an over-trust scenario is presented. Section 4.3 explains the evaluation under bi-directional trust changes. In section 4.4, the evaluation with continuous cooperative tasks is described.

4.1 Overview

We conduct three empirical studies to evaluate the proposed methods under several trust change scenarios with two types of cooperative tasks. All the evaluations are done with online experiments using a web-based drone simulator. Participants are recruited through crowdsourcing services.

Table 4.1: Three empirical studies.

Study	Evaluation	Scenario	Task
1	Evaluation with four types of TCCs	Over-trust scenario	Pothole inspection
2	Evaluation under bi-directional trust changes	ABA/BAB scenarios (A:under-trust, B:over-trust)	Pothole inspection
3	Evaluation with continuous cooperative tasks	ABA scenario	Drone navigation

Three Empirical Studies

The first empirical study described in section 4.2 is done with pothole inspection tasks to verify how the proposed method works in a simple over-trust scenario. Four types of TCCs are evaluated to compare the extent to which each TCC can change participants' choice behaviors.

The second empirical study described in section 4.3 is done with pothole inspection tasks to investigate the proposed method in bidirectional trust changes of ABA/BAB scenarios of under-trust (A) and over-trust (B).

The third empirical study described in section 4.4 is done with drone navigation tasks to examine how effective the proposed method would be in real-time applications such as autonomous driving.

Table 4.1 summarizes the three empirical studies.

Two Cooperative Tasks

Two types of cooperative tasks are prepared for the evaluations: a pothole inspection task and a drone navigation task. Table 4.2 compares the key characteristics of the two cooperative tasks.

A pothole inspection task is a visual search task to check if there were any holes or cracks in the road images from the drone. Each inspection task is executed *discretely* at every checkpoint when the drone reaches one, and they are *mutually independent*. This type of visual inspection can be categorized as a reconnaissance task, which is often used in the trust research literature. Applications such as baggage inspection

Table 4.2: Comparisons of cooperative tasks used in empirical studies.

	Pothole Inspection (in 1st and 2nd studies)	Drone Navigation (in 3rd study)
Task	Discrete Every check point	Continuous Every 0.12 seconds
State dependency	Mutually independent Road surface conditions	Dependent on prev. task Drone's positions and directions
Applications in same category	Baggage inspection Medical image diagnosis Product visual inspection	Autonomous driving (Lv4) Supervised unmanned vehicles Telepresence robot navigation

systems, medical image diagnosis, and product visual inspection systems are in the same category.

A drone navigation task is a real-time control task of navigating a drone to reach a goal along a predefined course. This task is *continuously* executed by cooperative activities between auto-pilot operated by a 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. Practical applications such as autonomous driving (SAE level 4), supervised unmanned vehicles, and telepresence robot navigation fall into the same category as this task.

Four types of TCCs

We designed four different types of TCCs: a visual TCC, audio TCC, verbal TCC, and anthro. TCC.

Cowell et.al [104] discussed the five non-verbal behaviors of an embodied conversational agent. Waytz et.al [105] demonstrated that anthropomorphism increases trust in an autonomous vehicle. Laughert et al. [106] examined three important objectives in effective warnings: attract attention, elicit knowledge, and enable compliance behavior. Based on these pieces of literature, we designed and evaluated the four types of TCCs (Fig 4.1) in our experiment.

The visual TCC was designed as a red warning sign in the shape of an upside-down

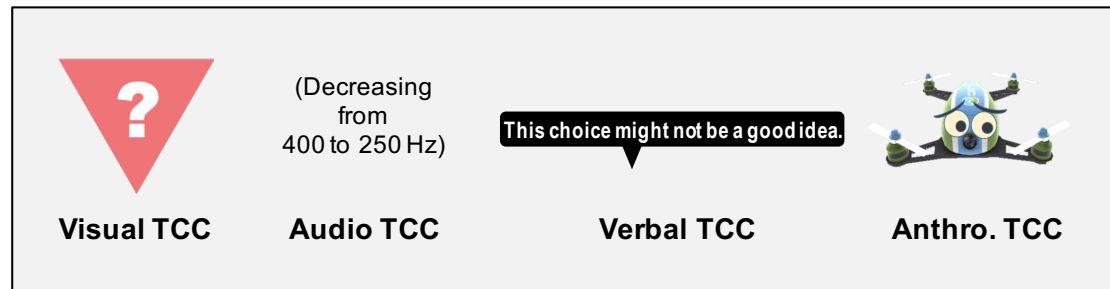


Figure 4.1: Four types of TCCs.

triangle, which is considered to be one of the most common alerting signs according to [107, 106]. The audio TCC uses a sound with a frequency that decreases from 400 Hz to 250 Hz as a negative message [103]. Verbal TCC is a tooltip balloon with the warning message “This choice might not be a good idea.” The anthropomorphic TCC is an animated drone image with a cartoon-like unhappy face. All of the four TCCs were designed not to be messages with specific system information, but to be notifications.

4.2 Evaluation with an Over-trust Scenario

This section presents our first empirical study to verify how the proposed method works in a simple over-trust scenario. Four different types of trust calibration cues (TCCs) are evaluated.

4.2.1 Introduction

We conducted an online experiment with a web-based drone simulator to evaluate the our method’s effectiveness in an over-trust scenario. The participants of the experiment performed a pothole inspection task [108] to check if there were any holes or cracks in the road images from the drone. Participants chose to use the drone’s automatic inspection or to check the road image manually. The proposed framework judged the trust calibration status by observing the participants’ choice behavior and presented TCCs when over-trust was detected. We measured behavioral changes to see if our adaptive method could effectively restore an appropriate status of trust calibration.

We expected users to change their choice behavior if TCCs were adaptively

presented when the framework detected inappropriate trust calibration. If our method could effectively mitigate the over-trust or under-trust, the following are hypothesized:

[H1-0] the manual choice rates increase in cases of over-trust or decrease in cases of under-trust.

[H1-1] the users with TCCs perform better than the users without TCCs.

4.2.2 Method

All studies were carried out in accordance with 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 Ministry of Health, Labor, and Welfare in Japan with written informed consent from all participants. All participants gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the ethics committee of the National Institute of Informatics.

Participants

We recruited participants online through a crowdsourcing service provided by Macromill, Inc. Regarding online experiments in general, Crump et al. [109] 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.

194 participants joined the experiment online. They were between 20 to 69 years old ($M = 44.35$, $SD = 14.10$). 96 participants were male and 98 were female.

Apparatus and Materials

We developed a 3D drone simulator based on an open-source JavaScript WebGL library CesiumJS [110] and the Bing Map API [111]. A screenshot of the simulator running on a Chrome browser is shown in Fig 4.2.¹

¹All map images in this section 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 4.2: Online drone simulator.

Operating the simulator was relatively easy, with two cursor keys for controlling the direction of the drone and mouse buttons for making choices.

Pothole Inspection Tasks

A route with 30 checkpoints (CKPs) was defined in the simulated environment. Each CKP was located in the center of a rectangular area that was to be inspected to see if there were any potholes in it. Out of the 30 CKPs, 10 had potholes in the corresponding areas while the other 20 did not. CKPs on the route were shown as small yellow circles on the screen. When the drone came close enough to one of the CKPs on the route, a message popped up (Fig 4.3) in which the drone asked the participants to make a choice.

The indicator at the bottom left area of the screen always showed the reliability of the automatic pothole inspection (Fig 4.4). This continuous display helped to increase the system transparency in terms of the reliability.

If the participants selected the “Auto” button, an automatic-inspection result was shown for three seconds with a road image of the area around the CKP. This feedback

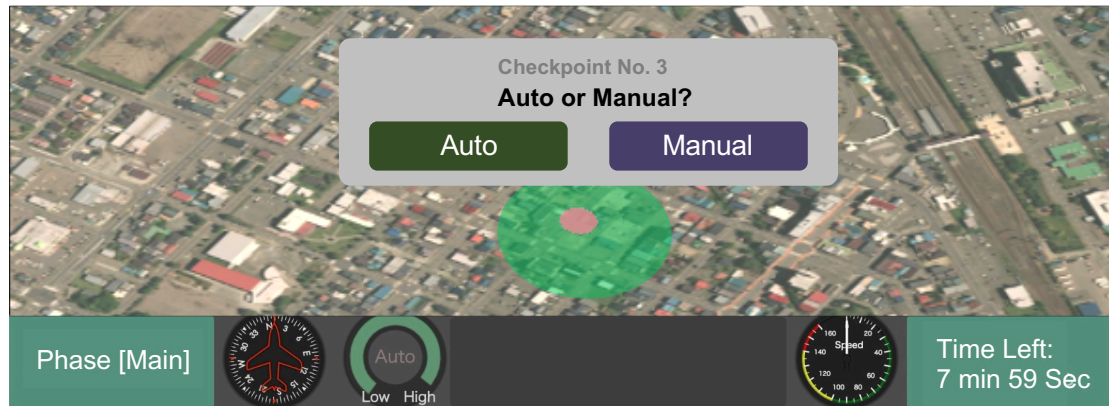


Figure 4.3: Popup message asking the participants for choice.

information helped the participants understand how well the automatic inspection performed, thereby increasing the system transparency [8, 79]. If the “Manual” button was selected, a road image was displayed, and the participants had to make a pothole report manually. Popup windows of both cases are as shown in Fig 4.5. Potholes were artificially rendered as irregular shapes in a dark brown color on a road image in the popup window.

Four TCCs

We evaluated the four different types of TCCs (see Figure 4.1) in a scenario that was designed to incur the user’s over-trust. TCCs were presented to the participants when the framework detected the over-trust status: the audio TCC was played once, and other TCCs were displayed on the screen close to the “Auto” button for two seconds.

Manipulation

The parameter P_A was manipulated to evaluate our method in an over-trust scenario. The performance of the automatic pothole inspection was configured based on signal detection theory (SDT) [112]. The SDT describes the detection of signals in noisy environments. The noise and the signal are represented as two overlapping density distributions. The distance between the two curves represents the sensitivity d' of the system.



Figure 4.4: Reliability Indicator at the bottom left area of the screen. Showing a higher reliability (left) and a deteriorated reliability (right).

In this experiment, the underlying base rate of potholes was 0.3. Under good weather conditions, P_A and the corresponding sensitivity d' were 90% and 1.8 respectively, indicating a pretty good discriminating ability of the system. Under bad weather conditions, P_A dropped to 50% and the corresponding sensitivity d' became 0.1, meaning the reliability of the automatic pothole inspection had greatly deteriorated. Fig 4.6 illustrates the manipulation of P_A and its relationship with \hat{P}_A and P_H .

Participants were expected to reach at least the 15th CKP in this experiment.

Assumption

The images of the potholes were carefully designed so that the average success rate of manual inspection would be more than 75%. Although machine image recognition has advanced remarkably with deep neural networks [113], several studies including [114, 115] demonstrated that human object recognition outperforms the top-performing deep neural networks under image degradation, such as Gaussian blur and additive Gaussian noise. Humans are better at generalizing across a wide variety of changes in an input image distribution, such as across different illumination conditions and weather types.

These findings could be applied to the estimation of the inequalities of P_H and P_A in the experiment because the pothole inspection became an image recognition



Figure 4.5: Popup windows of the pothole inspections. Automatic inspection result window (left) . Manual inspection window (right) . Both images contain potholes as dark brown spots in the upper road areas.

task with dark and foggy road images when the weather conditions turned worse. Therefore, we assumed that P_A would fluctuate more widely than P_H under changing weather conditions. On the basis of this assumption, we calculated the inequality relationship between P_A and P_H in the experiment; the inequality $P_A > P_H$ was true during the good weather period and false during the bad one.

Procedure

Participants were randomly assigned to one of five groups with the corresponding TCCs: NoTCC group (without TCCs), visual group (with the visual sign TCC), audio group (with the audio TCC), verbal group (with the verbal TCC), and anthro. group (with the anthropomorphic TCC). The NoTCC group was the control group in this experiment.

The experiment was completed online in three phases. In **the instruction phase**, the participants were given instructions stating the goal of the experiment was to inspect 22 CKPs out of 30 CKPs on a test route within a time limit. The participants

Algorithm A1 : Over-trust detection

Initialize:Total number of check points(CKPs): M = the number of CKPs.;Over-trust flag list: $OT[1], \dots, OT[M]$ are initialized with zero;The number of current CKP: $i \leftarrow 1$;**while** $i \leq M$ **and** not time-over **do** **if** the drone reached a CKP **then** **if** choice behavior is AUTO **and** $P_{man} > P_{auto}$ **then** $OT[i] \leftarrow 1$; **if** $i \geq 3$ **and** $(OT[i - 2] + OT[i - 1]) \geq 1$ **then**

Over-trust is detected and TCC is presented to the user;

end if **end if** $i \leftarrow i + 1$; **end if****end while**

learned they could inspect CKPs by checking the road image manually or by relying on the drone's automatic pothole inspection capability. They were told that the average success rate of manual pothole inspection was 75% so that they could adjust their initial self-confidence \hat{P}_{man} accordingly. They also learned that the reliability of the automatic pothole inspection was very high, although it could fluctuate depending on the conditions of the weather and sunshine. At the end of the instructions, the participants were guided to adjust the sound volume level by listening to a 400 Hz beep sound. Next, in **the training phase**, the participants started to fly the simulated drone in the training mode. They learned how to operate the drone and how to inspect the CKPs with on-screen guides. When the first three CKPs were inspected, the training mode was finished and **the main phase** was started. The reliability of the drone's pothole inspection P_A was artificially manipulated by changing the conditions of the weather and sunshine in the simulated environment. Initially, the weather was good, and P_A was set to 90%. The fine weather continued until the drone visited six CKPs in the main phase. This period of six CKPs was intended for the participants to calibrate their trust toward the drone with a higher reliability of automatic inspection under the good weather conditions. Immediately after the 6th CKP was inspected, sounds of a thunderstorm began. The visibility of the field also became very low, and the P_A was

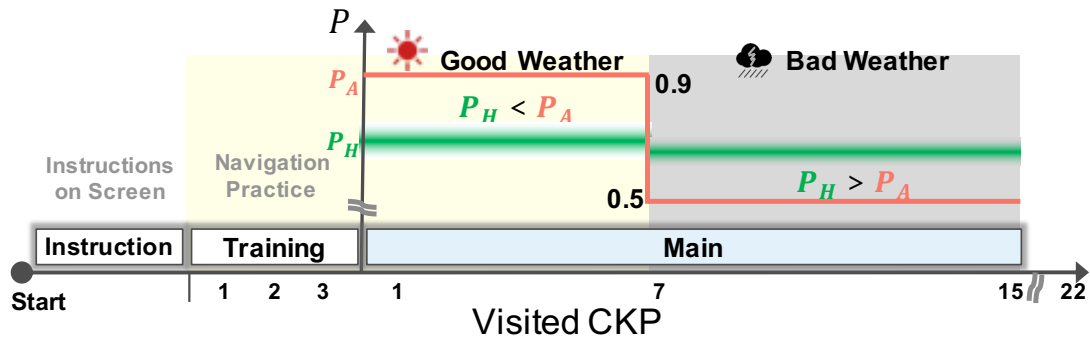


Figure 4.6: Relationship between the three parameters under changing weather conditions

decreased from 90% to 50%, which changed the sign of $P_H - P_A$. During this period, the participants were expected to over-trust the drone due to carry-over from the previous weather condition. The proposed framework was evaluated in this period. When the participants clicked the “Auto” button for automatic inspection and the framework detected the over-trust status using the over-trust detection algorithm [A1](#), the corresponding TCC was presented at the timing right after the button was selected. Note that a simple moving average of three CKP windows was used in the algorithm. The following information was available to the participants during the experiment:

- Simulated weather information with visibility changes, brightness changes, and the sounds of a thunderstorm.
- A reliability indicator to continuously show the reliability of the automatic inspection.
- An enlarged photo images of each CKP.
- The result of the drone’s automatic inspection when AUTO was chosen.
- A TCC when the proposed framework detected over-trust.

The experiment ended if the 22nd CKP was inspected or the time limit was reached. We established a time limit of 8.5 minutes (510 seconds) based on pre-trials with this test route, and we expected a single automatic inspection to take 10 seconds, one manual inspection to take 15 seconds, and reaching the next CKP to take 10 seconds.

Dependent Measures

The dependent variables of interest in this experiment were a TCC presentation rate (hereinafter, called TCC rate), a manual choice rate (hereinafter, called manual rate), sensitivity d' , and accuracy of the task results. TCC rates mean how often the proposed framework detected over-trust. Changes in manual rates indicate how the participants changed their behaviors as a result of trust calibration. Both sensitivity d' and accuracy indicate the performance of the human-AI cooperation.

All keyboard inputs and mouse clicks were recorded and used to calculate these variables.

4.2.3 Results

One hundred sixteen participants successfully inspected 15 CKPs or more within the time limit. Seventy eight participants unintentionally moved the drone far from the area where the CKPs were located, and they failed to complete the tasks within the time limit. As for the successful participants, their ages ranged from 20 to 69 years old ($M = 43.25, SD = 14.01$), 66 participants were male and 50 were females. 28 were in the NoTCC group, 18 in the visual group, 22 in the audio group, 29 in the verbal group, and 19 in the anthro. group. They inspected the total of 1,740 CKPs from the 1st CKP to the 15th CKP, and the results of 1,282 inspections were correct, making the correct answer rate 0.74. Automatic inspection was selected 1,236 times (the choice rate = 0.71). The participants did the manual inspection 504 times (the choice rate = 0.29). Table 4.3 shows the TCC rates at each CKP. Note that TCCs were not presented in the period from the 7th CKP to the 9th CKP, since the sliding window of three CKPs was used in the detection algorithm. Means and standard errors of the other dependent measures can be found in Table 4.4. Hereinafter, we call the period from the 1st visited CKP to 6th CKP “the good weather period” and the period with possible TCC presentations from the 10th CKP to 15th CKP “the bad weather period”.

TCC Rates

TCC rates in the verbal TCC group were higher in the early part of the period and gradually decreased. This indicates that over-trust decreased during this period. The

Table 4.3: Means of TCC rates at each CKP.

group	CKP9	CKP10	CKP11	CKP12	CKP13	CKP14	CKP15
Visual	0.78 (0.10)	0.67 (0.11)	0.56 (0.12)	0.67 (0.11)	0.67 (0.11)	0.50 (0.12)	0.56 (0.12)
Audio	0.55 (0.11)	0.64 (0.10)	0.45 (0.11)	0.50 (0.11)	0.50 (0.11)	0.50 (0.11)	0.50 (0.11)
Verbal	0.48 (0.09)	0.45 (0.09)	0.28 (0.08)	0.31 (0.09)	0.34 (0.09)	0.38 (0.09)	0.07 (0.05)
Anthro.	0.53 (0.11)	0.47 (0.11)	0.37 (0.11)	0.53 (0.11)	0.47 (0.11)	0.74 (0.10)	0.63 (0.11)

Standard errors in parentheses.

Table 4.4: Means of the other dependent measures

Group	Manual rate		Sensitivity d'		Accuracy	
	GW	BW	GW	BW	GW	BW
NoTCC	0.15 (0.04)	0.22 (0.06)	1.36 (0.09)	0.38 (0.14)	0.86 (0.04)	0.60 (0.04)
Visual	0.11 (0.06)	0.37 (0.09)	1.35 (0.13)	0.52 (0.15)	0.87 (0.05)	0.62 (0.04)
Audio	0.20 (0.06)	0.42 (0.08)	1.62 (0.10)	0.61 (0.15)	0.94 (0.03)	0.65 (0.04)
Verbal	0.17 (0.05)	0.63 (0.04)	1.39 (0.10)	0.92 (0.14)	0.87 (0.04)	0.72 (0.04)
Anthro.	0.20 (0.06)	0.37 (0.07)	1.40 (0.15)	0.54 (0.15)	0.88 (0.05)	0.62 (0.04)

Standard errors in parentheses.

“GW” means the good weather period and “BW” means the bad weather period.

visual and audio TCC groups also showed a similar trend, while TCC rates in the anthro. TCC group did not follow the decreasing trend.

Manual Rates

TCCs were presented multiple times per participant in most cases. The effects of presenting TCCs might be accumulated and did not always appear immediately after presentation. In the current study, we evaluated the TCC effects by comparing the six-CKP mean values of the manual rates both for the good and bad weather periods, so that we could also capture the accumulated effects in each period. We conducted a

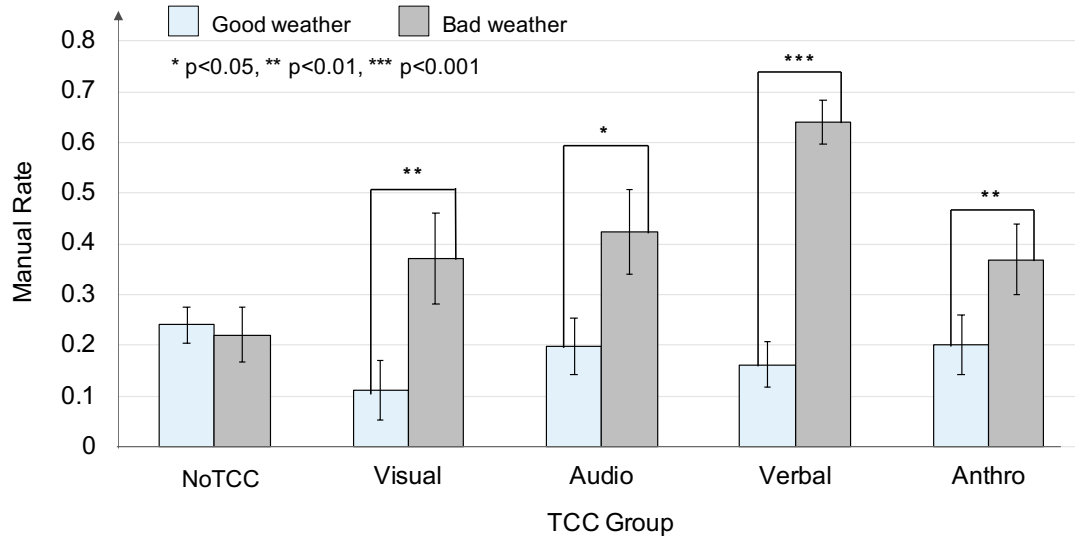


Figure 4.7: Manual rates over time

two factor mixed ANOVA with the TCC groups (NoTCC, visual, audio, verbal, and anthro.) as between subjects and CKP periods (the good weather period and the bad weather period) as within subjects. The analysis revealed a significant main effect for the CKP periods [$F(1, 111) = 51.69, p < 0.01, \eta_p^2 = 0.32$]. The participants changed their choice behavior as the weather conditions deteriorated. A significant interaction was found between the two factors [$F(4, 111) = 4.86, p < 0.01, \eta_p^2 = 0.15$].

In the good weather period, there was no simple effect for the TCC groups, meaning that the manual rates of each TCC group were not significantly different from each other [$F(4, 111) = 0.47, p = 0.76, \eta_p^2 = 0.02$]. The NoTCC group did not show a simple effect for the CKP periods [$F(1, 27) = 1.23, p = 0.28, \eta_p^2 = 0.04$] indicating the manual rates of the NoTCC group were not significantly different between the two CKP periods. In contrast with this, all of the groups with TCCs showed significantly higher manual rates in the bad weather period than in the good weather period [Visual TCC, $F(1, 17) = 9.20, p < 0.01, \eta_p^2 = 0.35$; Audio TCC, $F(1, 21) = 5.54, p = 0.03, \eta_p^2 = 0.21$; Verbal TCC, $F(1, 28) = 62.9, p < 0.001, \eta_p^2 = 0.69$; Anthro TCC, $F(1, 18) = 8.55, p < 0.01, \eta_p^2 = 0.32$]. Fig 4.7 shows how the manual rates changed over two CKP periods.

Holm–Bonferroni-adjusted post hoc comparisons were also conducted to investigate

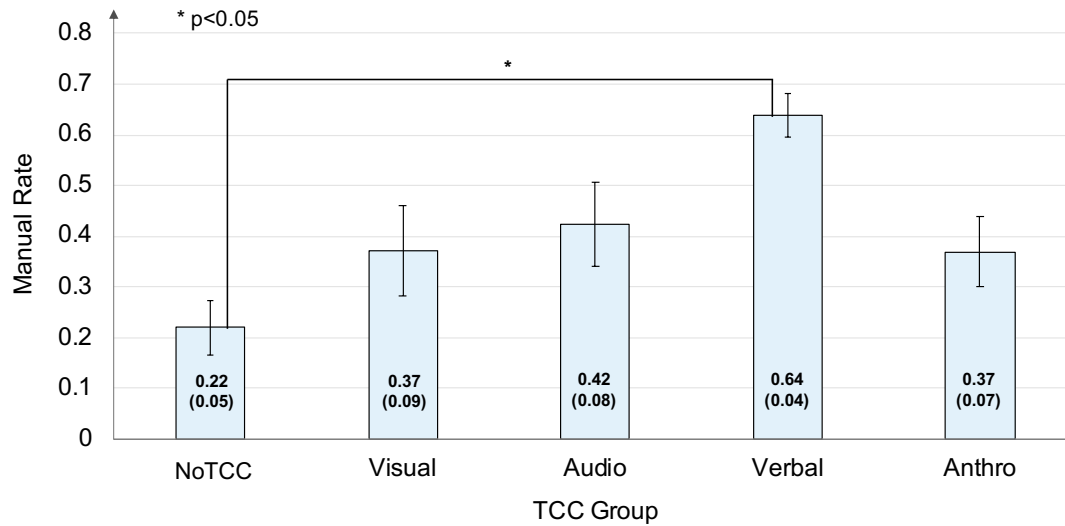


Figure 4.8: Manual rates for each TCC during the bad weather period

the effects of TCCs. For the bad weather period, the verbal group showed a significantly higher manual rate than both the NoTCC group ($t(111) = 4.77$, $adj.p < 0.01$) and the anthro. group ($t(111) = 2.74$, $adj.p = 0.04$). No other differences between the groups were found to be significant. Although the effectiveness among TCCs differ, these results support **H1-0**.

Performance : Sensitivity and Accuracy

The same ANOVA with the TCC groups and the CKP periods revealed that the sensitivity d' in the bad weather period was significantly lower than in the good weather period [$F(1, 111) = 107.22$, $p < 0.01$, $\eta_p^2 = 0.49$]. In the bad weather period, post-hoc comparisons indicated that the sensitivity d' of the verbal group, which was the highest (0.92 (SE 0.14) among all the groups, was significantly higher than the sensitivity d' of the NoTCC group [$t(111) = 2.97$, $adj.p = 0.04$, $Cohen's d = 0.75$]. In terms of sensitivity d' , hypothesis **H1-1** is supported.

Accuracy, the rate of the correct inspection, also significantly declined in the bad weather period [$F(1, 111) = 87.58$, $p < 0.01$, $\eta_p^2 = 0.46$], but there was no significant

Table 4.5: 3-CKP mean values of P_H

CKPs	Good weather period		Bad weather period		
	1 to 3	4 to 6	7 to 9	10 to 12	13 to 15
P_H	0.81 (0.05)	0.80 (0.06)	0.85 (0.04)	0.86 (0.03)	0.90 (0.03)

Standard errors in parentheses.

difference among the five groups [$F(4, 111) = 1.62, p = 0.18, \eta_p^2 = 0.06$]. Hypothesis **H1-1** is not supported regarding accuracy; however, the verbal group showed the highest mean value (0.72 (SE 0.04)), and other groups with TCCs also had better accuracy values (0.6 (SE 0.04)) than the NoTCC group.

Regarding the accuracy of the manual inspections which corresponds to P_H , Table 4.5 shows the 3-CKP mean values of P_H . Although the mean values of P_H slightly increased, a Welch's t-test indicated that there was no significant difference between P_H in the good weather period and in the bad one [$t(114) = -1.08, p = 0.28, \text{Cohen's } d = 0.16$]. This result indicates that P_H did not degrade under the change in weather conditions. One-sample t-tests showed that P_H in the good weather period was significantly smaller than $P_{A=0.90}$ [$\text{Mean} = 0.81, t(71) = -2.17, p = 0.03, \text{Cohen's } d = 0.26$] and that P_H in the bad weather period was significantly larger than $P_{A=0.50}$ [$\text{Mean} = 0.86, t(200) = 17.79, p < 0.01, \text{Cohen's } d = 1.25$].

4.2.4 Discussion

Effects of TCCs to Change the Participants' Behavior

The groups with visual, verbal, and anthro. TCCs showed significantly higher manual rates for the bad weather period than for the good weather period (Fig 4.7). The audio TCC that used an audio ASE that decreased in frequency seemed promising, as it could convey the low level of confidence in the system [103]. The result, however, showed that its manual rate was not significantly larger than the one of NoTCC group. One possible reason is that other sounds were also being played in the simulated environment, such as the drone flying and the thunderstorm, which may have reduced

the effect of this audio TCC. The effect of the anthro. TCC was not as large as we originally expected; it was significantly smaller than the effect of the verbal TCC (Fig 4.8). The manual rate of the visual TCC were also smaller than other TCCs. It was obvious that the participants recognized these visually impressive TCCs; however, the results suggested that just being salient on the screen was not enough for some of the participants to change their choice behavior [116, 106]. The verbal group had the highest manual rate (Fig 4.8). Only the verbal TCC referred to the purpose by using the word “choice,” while the other TCCs were implemented just as a caution or warning. Reading this word might have helped the participants proceed more easily to the latter stages of the trust calibration process.

Based on these results, our tentative guideline for designing TCCs is that TCCs should be reasonably noticeable in the task environment and should contain connotations that can link the user to the next possible actions in the collaborative task.

If the participants wanted to complete the scan tasks quickly, they could have used the automatic inspection, which was faster. However, the participants promptly increased the manual choices after recognizing TCCs despite the longer completion time. This result indicates that the possible automation bias caused by the difference in the task completion times did not critically impact the decision-making of the participants in the experiment.

These results indicated that adaptively presenting TCCs strongly affected the choice behavior of the participants who otherwise failed to find opportunities to change their tendency to rely on the automation.

Performance

The results of the manual accuracy were consistent with our assumption to estimate P_H and P_A in this experiment.

While the mean values of sensitivity d' in each group were dropped in the bad weather period, the verbal group showed a significantly higher sensitivity d' than the NoTCC group, and three other groups with TCCs also had better values than the NoTCC group. The results showed that all the groups with TCCs showed higher discriminating performance in the bad weather period than the NoTCC group. Though

there was no statistical difference in accuracy among the groups, all the groups with TCCs also showed better accuracy than the NoTCC group.

These results indicated that adaptively presenting TCCs promoted appropriate trust calibration leading to the better performance in the bad weather period.

Adaptive Method vs. Continuous Method

The manual rate of the NoTCC group did not significantly change over the two periods. When the participants were exposed to the bad weather, the weather change was made very noticeable with the screen visibility and the sound effects. The reliability indicator showed a big performance degradation of the system due to the poor visibility. Nevertheless, the participants of the NoTCC group continued to rely on the drone's automatic pothole inspection, which had less reliability than the actual manual success rate. Thus, the participants over-trusted the automatic inspection despite the system information indicating the reliability becoming worse. This result is not in line with the previous studies [82, 78, 20] that emphasized the effectiveness of the continuous trust calibration with system transparency. A possible explanation for the result could be made by discussing models for the trust process [117, 1, 21]. Miring et al. [118] defined a model with four stages: perception, understanding, prediction, and adaptation. Although the reliability indicator continuously displayed the deterioration of the reliability, the participants in the NoTCC group might not fully acquire the knowledge to move beyond the perception stage. They would have behaved differently if the experiment had continued longer enough for them to understand the relationship between the indicator change and the performance of the system. In contrast to this, the participants in other groups with TCCs could successfully reach the adaptation stage and change their behaviors in this experiment. We believe that the results demonstrated the effectiveness of the adaptive method. TCCs were given right after the behavior only if the participants were judged to over-trust, so that it would be easier for the participants to understand the implication of the cues and to move forward in the trust calibration process.

Applicability in Real-world Situations

Although providing a model to estimate the second inequalities of P_H and P_A in the proposed framework is beyond the scope of this study, we believe that they could be estimated as follows. P_A , which represents the reliability of an AI system, could be calculated with the sensor models and algorithms used to implement the AI system. P_H , which is a human capability index, could be estimated by using the parameters of a target task and environmental conditions. The results of the previous studies [114, 115] are such examples that provide a basis for estimating the second inequalities. 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 use a bottom-up approach that utilizes the data from task executions if an appropriate estimation model for P_H is not available. For example, trial operations can be performed to collect the necessary data to estimate P_H empirically. In practical situations, it is quite common for users of a system to practice how to operate the system in advance. The second inequalities could be estimated even in a real-time situation. The first inequalities in the proposed framework could be estimated by observing users' choice behaviors, without measuring \widehat{P}_A and \widehat{P}_{man} directly. Although a pop-up dialogue was used to observe the behaviors in the current experiment, continuous measurements of the behaviors could also be used with the proposed framework. For example, a driver's intention to use automatic driving could be inferred with a touch sensor on a steering wheel to check if the driver's hands are on the wheel. Similarly, a switch button to turn automation on and off at any time could provide necessary information on humans' reliance on the automation. The first inequality in the framework could be calculated with the information from these continuous methods that could work well with real-time tasks. Therefore, we believe that our proposed framework can be applied to real-time applications that require human-AI cooperation. Our third empirical study described in section 4.4 evaluates the proposed method with a real-time task.

Feedback Information

The feedback information given for each inspection was very important for the participants to make decisions. The pothole inspection task in the experiment is a remote sensing task, and it would be quite difficult for the system to know the

correct answer (ground truth) at the time of each inspection in practical situations because the only information available is the image data and the results of automatic recognition. Therefore the correct answer for each inspection was not presented to the participants in the experiment. The result of the automatic inspection was shown to the participants when they selected automatic inspection, not when they did the inspections manually. Although this was to simplify the conditions and focus on evaluating the effect of presenting TCCs, further study should consider possible combinations of feedback information and evaluate their effects.

In this experiment, we focused on evaluating an over-trust case, which often has more serious adverse effects in actual situations [14, 13]. The second empirical study described in section 4.3 evaluates with bi-directional trust change scenarios to evaluate both cases of over-trust and under-trust.

4.2.5 Conclusion

The overall results demonstrated that adaptively presenting TCCs strongly affected whether the choice behavior of the participants would change, while continuously presenting the reliability information did not help the participants change their bias to rely on the automation. The better task performances were also achieved with the behavior changes triggered by TCCs, whose presentation timing was decided by the proposed framework. Previous studies emphasized the importance of the system transparency for proper trust calibration. Our results indicated that they are not always sufficient to recover from over-trust, and our method of adaptive trust calibration significantly helped the participants change their behavior and recover from the over-trust. Among the four TCCs tested in the experiments, the verbal TCC had the strong effect in changing the user behavior, although other TCCs also showed the effectiveness to calibrate trust. Despite several limitations, this study has demonstrated the effectiveness of presenting cognitive cues at the time of over-trust.

4.3 Evaluation under Bi-directional Trust Changes

This section presents our second empirical study to evaluate the proposed method under the bi-directional changes of trust conditions.

4.3.1 Introduction

In order to evaluate our proposed method both in bidirectional trust changes, we defined ABA/BAB scenarios of under-trust(A) and over-trust(B) by manipulating the weather conditions. We conducted an on-line experiment with the same drone simulator and the same common task with the scenarios as the first empirical study described in the previous section 4.2.

We expected users to change their choice behavior if TCCs were adaptively presented when the framework detected inappropriate trust calibration. If our method could effectively mitigate the over-trust or under-trust, the following are hypothesized:

[H2-0] the manual choice rates increase if TCCs are presented in cases of over-trust or decrease if TCCs are presented in cases of under-trust.

[H2-1] the users with TCCs perform better and more robustly than the users without TCCs.

[H2-2] adaptively presenting TCCs could trigger the trust calibration process more effectively than continuously maintaining system transparency.

4.3.2 Method

Verbal Cue as a TCC

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 4.9. 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).

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.

The purpose of the main experiment was to evaluate our framework for both bidirectional environmental changes. We defined the ABA/BAB scenarios of under-trust



Figure 4.9: Verbal TCC

(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) [112]. SDT describes the detection of signals in noisy environments. Noise and signals are represented as two overlapping density distributions. The distance between the two curves 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, meaning that the reliability of the automatic pothole inspection had greatly deteriorated.

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.

Procedures

The online experiment started with an **instruction phase**. The participants were given an instruction stating that the goal of the experiment was to inspect 24 CKPs within 20 minutes. They were told 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 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 learned 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 4.11 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. After the experiment, the participants were asked to fill out a post-experiment questionnaire.

The algorithm A2 "Adaptive Trust Calibration" based on the proposed method was applied in the experiment. This algorithm is essentially the same as the one used in the first empirical study except that this includes the part to detect the under-trust. A simple moving average of three CKPs was used in the algorithm to capture the

Algorithm A2 : Adaptive trust calibration

Initialize:Total number of check points(CKPs): M = the number of CKPs.;Over-trust flag list: $OT[1], \dots, OT[M]$ are initialized with zero;Under-trust flag list: $UT[1], \dots, UT[M]$ are initialized with zero;The number of current CKP: $i \leftarrow 1$;**while** $i \leq M$ **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** $i \geq 3$ **and** $(OT[i-2] + OT[i-1]) \geq 1$ **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** $i \geq 3$ **and** $(OU[i-2] + OU[i-1]) \geq 1$ **then**

Under-trust is detected and TCC is presented to the user;

end if **end if** $i \leftarrow i + 1$; **end if****end while**

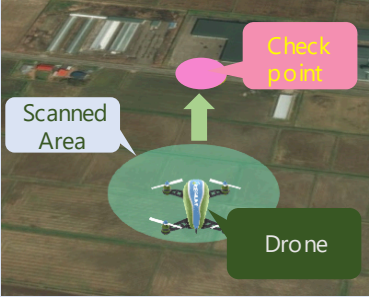
participants' behavior changes in each condition with eight CKPs.

Instruction

Please find the potholes on the road with the drone.

- Your goal is to inspect **24 checkpoints** displayed as yellow/red circles.
- The time limit is **20 minutes**.
- Your score will be decided by the completion time and the correctness.

Operation Guide




Drone Operation

Left ← → Right
Cursor Keys

Choices to Make


Auto Manual
I found a pothole No pothole

Mouse 

- This drone has an automatic pothole inspection function. The reliability is almost perfect, **close to 100%**, but it may change due to the weather conditions.
- The average correct rate for manual inspection would be around **75%**.


Information area at the bottom of the screen

Phase [Main]



Direction


Auto



Low High

Reliability indicator of the drone's automatic pothole inspection function

Speed



Speed

Time left:

7 min 59 sec

Figure 4.10: The instruction screen

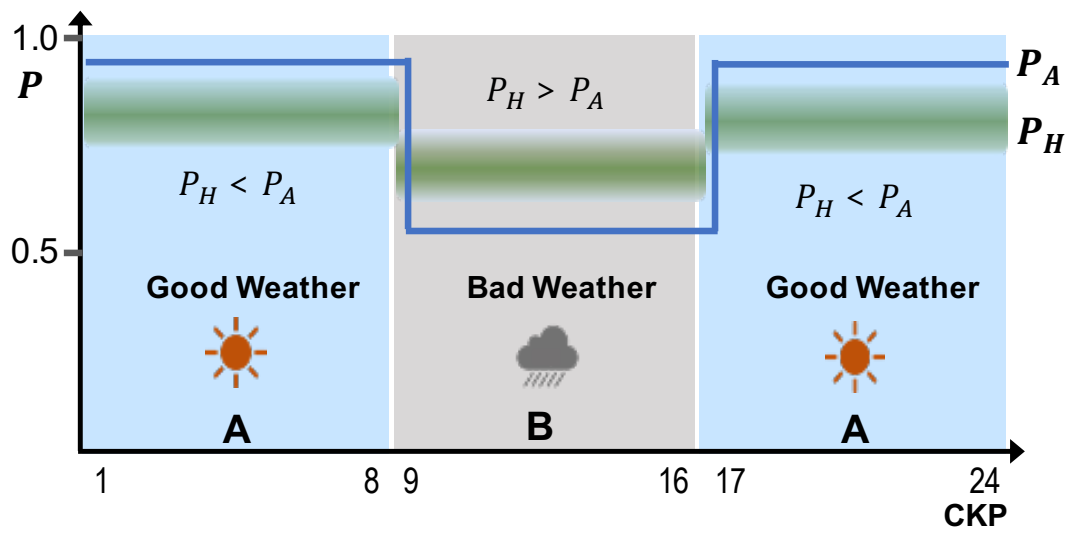


Figure 4.11: P_A and P_H in the ABA scenario

Assumption

The same assumptions are made as in the first empirical study described in the previous section 4.2. As the pothole inspection tasks are mainly image recognition tasks, we assumed that P_A would fluctuate more widely than P_H under changing weather conditions. On the basis of this assumption, we calculated the inequality relationship between P_A and P_H in the experiment; the inequality $P_A > P_H$ was 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 assumptions on P_H . 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 in accordance with 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.

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 method 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

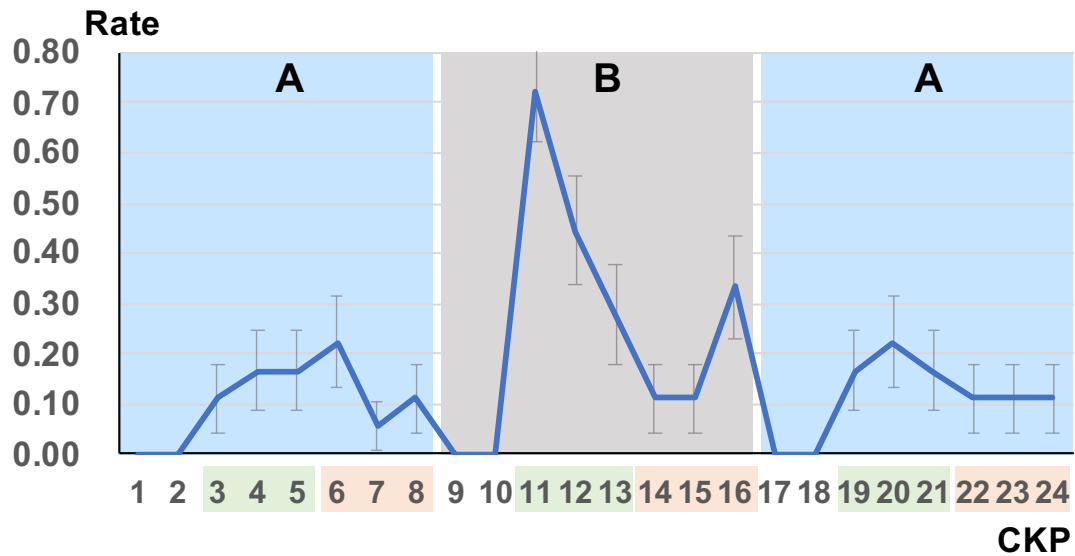


Figure 4.12: TCC rates of TCC-ABA group

automatic inspection and therefore indicating their trust status. The sensitivity d' demonstrates the performance of human-AI collaborative tasks.

4.3.3 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.

TCC Rates

Figure 4.12 and Figure 4.13 illustrate the TCC rates at each CKP of the TCC-ABA group and the TCC-BAB group. Table 4.6 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.

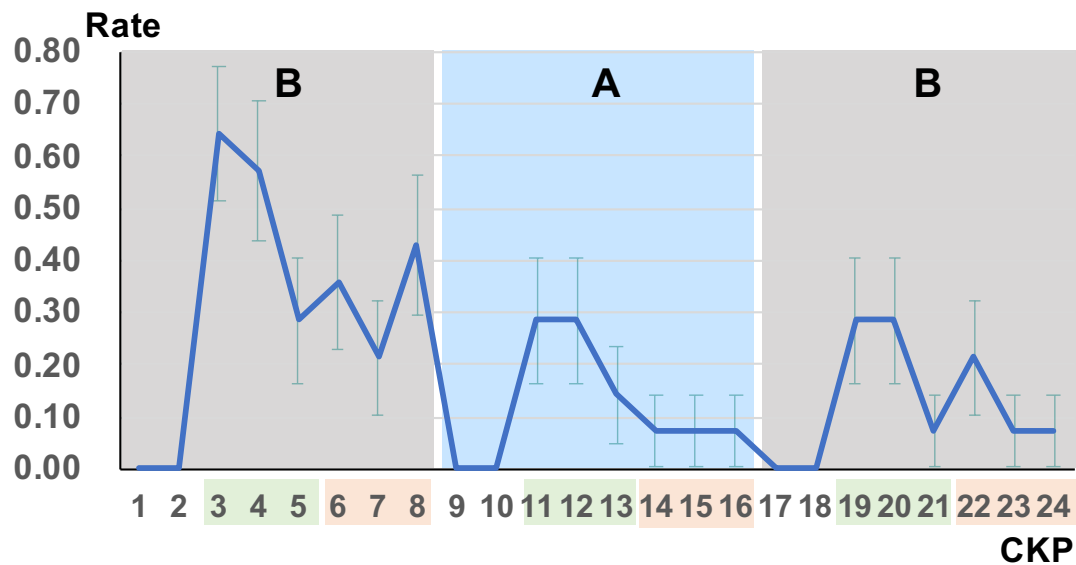


Figure 4.13: TCC rates of TCC-BAB group

ABA groups: The mean of the TCC rates from CKP 3 to 5 [hereinafter referred to as TR (3-5)] for the first A condition (C1) was low at 0.15 and slightly decreased to 0.13 for TR (6-8). We did a paired t-test that revealed that there was no significant difference between TR (3-5) and TR (6-8). For the B condition (C2), the TCC rate went up to the maximum at CKP 11. TR (11-13) was 0.48 and quickly decreased after that. A paired t-test showed that TR (14-16) was significantly lower than TR (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 at the almost same level as the first A condition. The difference between TR (19-21) and TR (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 revealed that TR (6-8) was significantly lower than TR (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 TR (14-16) significantly decreased from TR (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 TR (19-21) and TR (22-24), although TCC rates slightly decreased during the condition.

Table 4.6: 3-CKP means of TCC rates in each condition

CKPs	C1			C2			C3		
	3-5	<i>p</i>	6-8	11-13	<i>p</i>	14-16	19-21	<i>p</i>	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	0.50 (0.13)	*	0.33 (0.13)	0.24 (0.11)	*	0.07 (0.07)	0.21 (0.11)		0.12 (0.09)

* $p < 0.05$, ** $p < 0.01$

Table 4.7: Means of the manual rates and the sensitivity d'

Condition	Manual rate			Sensitivity d'		
	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)

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.

The Manual Rate

The change in manual rates indicates how the participants changed their trust in the automatic inspection. Building trust is an accumulating process [1], 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 method 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 4.7 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 4.14 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 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 [$t(20) = 4.47, adjusted.p < 0.01$].

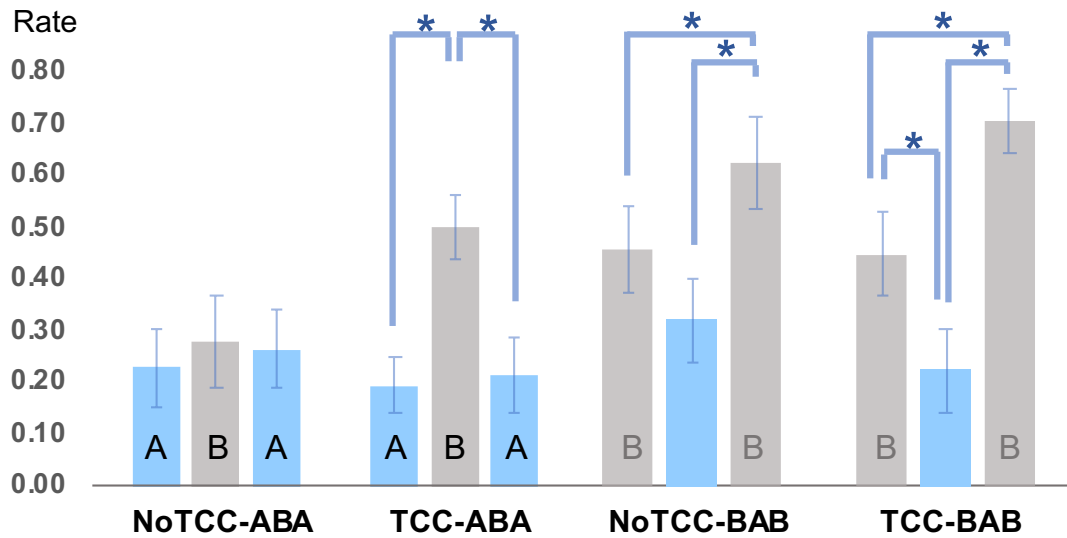


Figure 4.14: Manual rates

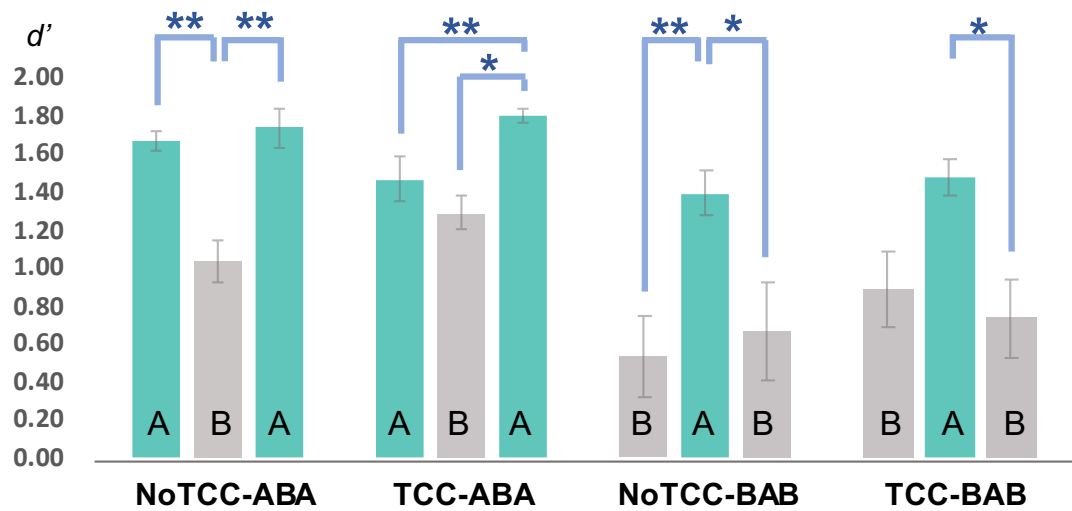
* $p < 0.05$

Performance

We conducted the same one-way ANOVA with the sensitivity d' of each group. Figure 4.15 illustrates the sensitivity d' for each condition of each groups.

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

Figure 4.15: Sensitivity d'

* $p < 0.05$, ** $p < 0.01$

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* $= 0.02$].

4.3.4 Discussion

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 capability of the drone under the good weather conditions, probably on the basis of their knowledge on the reliability of the automatic inspection 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 under-trust and suggest that trust calibration done to mitigate under-trust was successfully promoted by the proposed method. 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 **H2-0**. Regarding the performance, the results of the sensitivity d' confirm hypothesis **H2-1**, 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 properly calibrate the trust. This results support the result observed in the over-trust scenario of the first empirical study. 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 **H2-2**. TCCs were given right after the behavior only if the participants were judged to be in a state of over-trust or under-trust, so it would be easier for the participants to understand the implication 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[20] or perfect automation schema [78].

Several limitations of our study suggest the need for further experiments. In the proposed detection algorithm, a binary decision is made with a simple moving average

value of three CKPs. 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. The current study mainly dealt with dynamic learned trust [2]. Future studies should investigate other factors of trust, such as dispositional trust [1] and situational trust to gain a deeper understanding of trust calibration.

We used a pothole inspection task in the experiment, which is often categorized as a reconnaissance task in the trust research literature [50]. This type of tasks are performed independently and discontinuously.

4.3.5 Conclusion

The results of the experiment found clear support for the effectiveness of adaptively presenting a simple cue in changing the participants' reliance on autonomous systems in both cases of over-trust and under-trust. Even in the conditions where the continuous system transparency did not work well to help the participants properly calibrate their trust, the proposed method could assist the participants change their choice behaviors. Despite several limitations, the current study has demonstrated that the proposed method successfully promoted trust calibration in the case of both over-trust and under-trust caused by environmental changes. With the results of the first two empirical studies, we concluded that the proposed method could be a first step to answer the two research questions in section 1.2.

4.4 Evaluation with Continuous Cooperative Tasks

This section presents the third empirical study to evaluate the proposed method with continuous cooperative tasks.

4.4.1 Introduction

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 two empirical evaluations, the participants' selection decisions and operations must be made quickly enough to control the drone smoothly. The goal of the study was to evaluate the effectiveness of the proposed method in a real-time application environment.

4.4.2 Method

To apply the proposed method to human-AI cooperation with continuous cooperative tasks, we modified the behavior measurement to capture users' choice behaviors.

The first terms of (3.1) and (3.2) can be evaluated by observing the user's behaviors. As describe before, the reliance behaviors of a user can be explained by the user's perception of the reliability of a system and the user's own capability. When a user decides to rely on a system, it is reasonable to say that this behavior indicates $\widehat{P}_A > P_H$. If the user decides to do a task manually, it means $\widehat{P}_A < P_H$. If a cooperative task is not continuous, the evaluations are self-explanatory since each task involves a single choice behavior.

Modified Behavior Measurement

In the case of a continuous cooperative task where both the systems and the users can take over the control at any time during the task execution, the first terms can be evaluated as follows.

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 ($t \geq w$).

$$B_t = \frac{1}{w} \sum_{i=t-w}^t b_i \quad (4.1)$$

, 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 (3.1) and (3.2), P_A could be calculated with the sensor models



Figure 4.16: Online semi-autonomous drone simulator

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.

Apparatus and Materials

We added an auto-pilot function to the drone simulator used in the previous experiments. Figure 4.16 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 4.17) with the exactly 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



Figure 4.17: The first part of the course

autonomous navigation. This type of control is hereinafter 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 4.18), 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 [119], 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.

The verbal TCC was presented in front of the drone (see Figure 4.19) 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.

Participants and Scenarios

A total of 36 on-line participants (30 male, 6 female) were recruited a cloud-sourcing service provided by Yahoo! Japan. Participants were randomly assigned to one of two

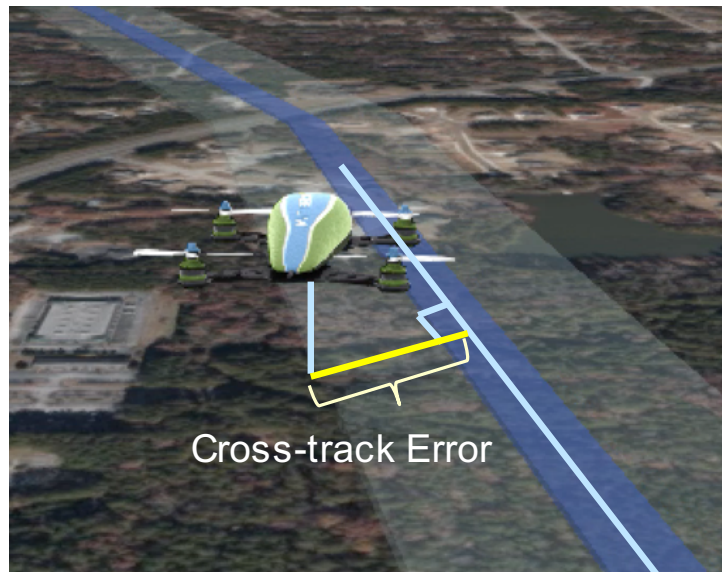


Figure 4.18: Cross-track error

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 ($M = 46.6$, $SD = 11.4$). They were recruited through a cloud-sourcing service provided by Yahoo! Japan.

We defined two scenarios of under-trust (A) and over-trust (B) by manipulating the reliability of the auto-pilot. In the A condition, good weather conditions were simulated. The screen brightness was 100%, and there were no sound effects except for

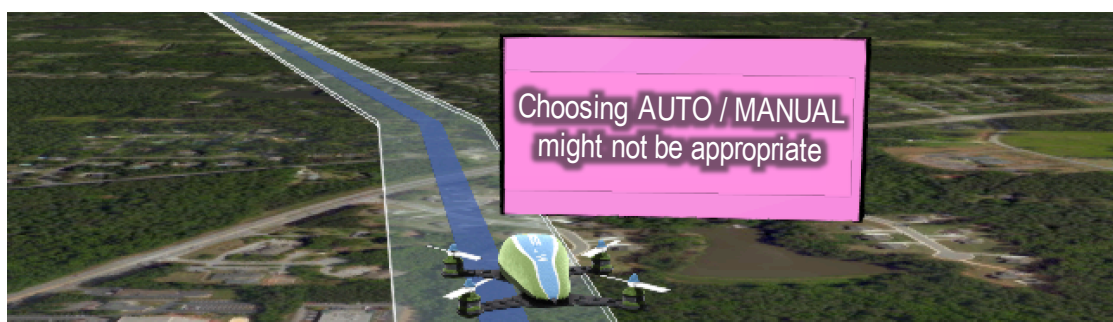


Figure 4.19: Verbal TCC

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 A1 condition and the A2 conditions respectively, which means the drone with auto-pilot flew accurately along the course. In the B condition, a thunder storm 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.

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.

Algorithm A3 : Adaptive trust calibration with continuous cooperative task

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

Evaluation of the Framework in the Experiment

The algorithm A3 “Adaptive Trust Calibration with Continuous Cooperative Task” based on the proposed method 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.

First terms. The moving average of the participants behavior were calculated at each timer-event. The window size was 12 seconds, which was suitable for capturing the changes in trajectory for the prepared course. **Second terms.** Although providing a general estimation model of P_H is beyond the scope of this paper, we estimated the second terms under the conditions of the current experiment in the basis of the robustness of human capability compared with that of recognition algorithms. 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 still remains challenging tasks [90]. Moreover, the work of [114] 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. 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 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.

Dependent Variables and Hypotheses

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 method 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.

If our method can effectively mitigate over-trust or under-trust, the following are hypothesized:

[H3-1] the manual-pilot rates increase if TCCs are presented in cases of over-trust or decrease if TCCs are presented in cases of under-trust.

[H3-2] the users with TCCs perform better than the users without TCCs.

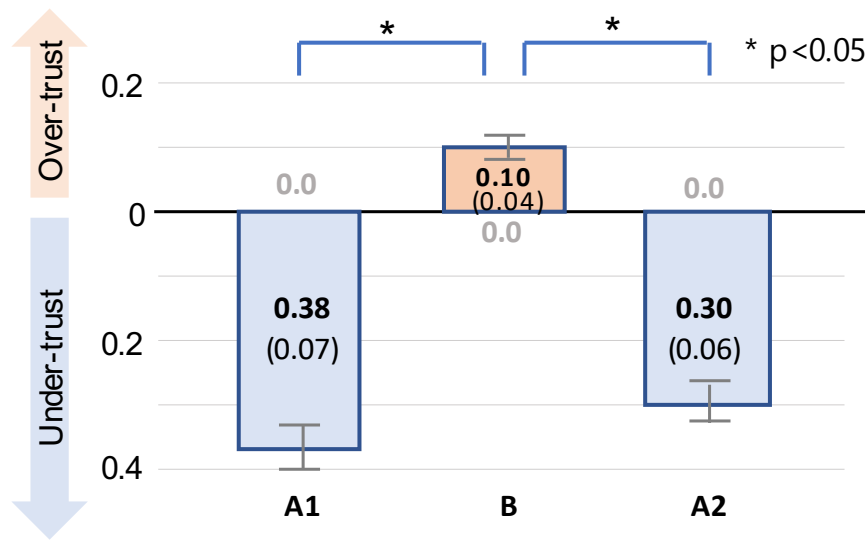


Figure 4.20: TCC rates

[H3-3] adaptively presenting TCCs could trigger the trust calibration process more effectively than continuously maintaining system transparency in a conventional way.

4.4.3 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 minutes YY seconds.

TCC Rates

Figure 4.20 illustrates TCC rates for each condition 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

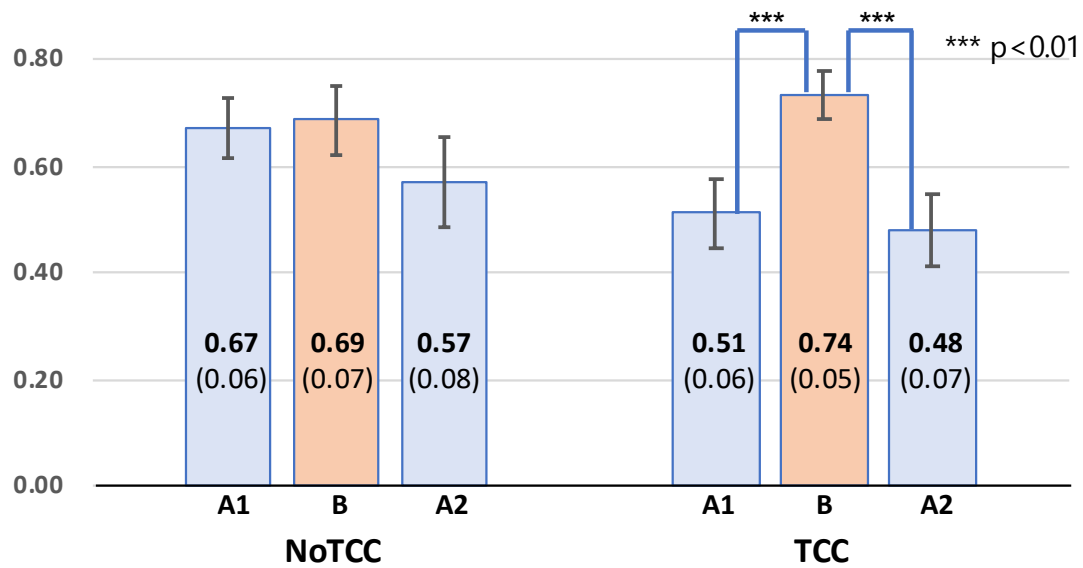


Figure 4.21: Manual-pilot rates

condition [$t(14) = 2.17$, $adj.p = 0.048$].

Manual-pilot Rates

Figure 4.21 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 the 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 [$t(14) = 5.72$, $adj.p < 0.001$].

Cross-Track Errors

Figure 4.22 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

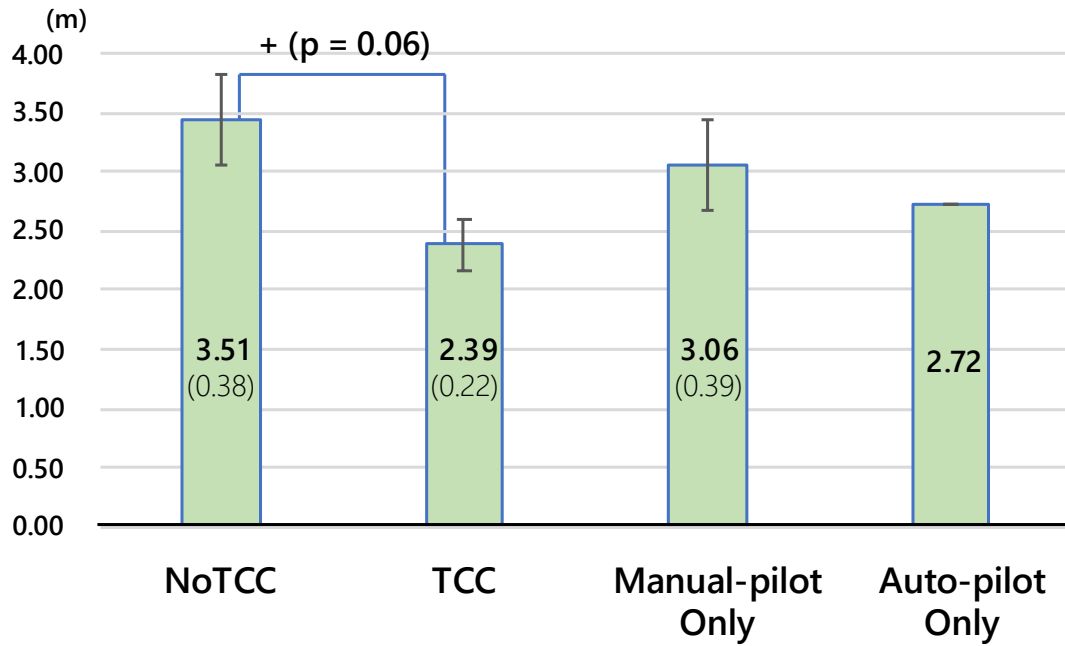


Figure 4.22: Cross-track errors

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 show 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 [$t(25.3) = 2.49$, $adj.p = 0.06$, $Cohen's d = 2.77$].

4.4.4 Discussion

The results of the TCC rates indicate that the proposed framework detected the under-trust and presented TCCs in the A1 and A2 conditions. The framework also detected the over-trust and presented TCCs in the B condition. The manual-pilot rates significantly increased from the A1 condition to the B condition, and significantly decreased from the B condition to A2 condition. As there were no such changes among the conditions observed in the NoTCC group, the hypothesis **H3-1** was supported. Note that the differences in the TCC rates among the ABA conditions indicate that the

task setting used in the experiment induced under-trust more than over-trust.

Although the difference was close to significant, the mean of cross-track errors for the TCC groups was smaller than that for the NoTCC group. Therefore, we consider that the hypothesis **H3-2** was partially confirmed. The TCCs worked as triggers to promote trust calibration and helped the participants improve the task performances by changing their choice behaviors.

In the NoTCC group, the manual-pilot rates were high even when the indicator showed high reliability of the auto-pilot in the A1 and A2 conditions. The strong tendency of under-trust was observed. Possible interpretation would be that the participants tended to intervene the auto-pilot when they noticed the drone was not heading along the course direction, even if the indicator displayed a high reliability of the auto-pilot. According to the post-experiment questionnaire results, forty percent of the participants answered that they selected the manual-pilot when they thought the drone was about to go out of the course. This preventive action taken by the participants suggested that the drone's behaviors had a greater impact on the participants' trust than the reliability indicator. In the TCC group, TCCs adaptively presented at the time of over-trust/under-trust was able to change the participants' behaviors. Therefore, these results supported the hypothesis **H3-3**.

We used a verbal TCC in this experiment, and its calibration effect was milder than in the first two experiments. This result may be due to a mental workload of reading the text message of verbal TCC during the real-time navigation task. The effects of visual TCC or audio TCC could be greater as they are more intuitive or with a different modality.

Several limitations in the current study suggest the necessity of the further research. The system transparency was realized in a rather simple way with the reliability indicator, the further study should consider how to improve the system transparency with the information on the intent, plans and processes of the autonomous systems [83]. The single task setting was examined in the current study, however, the benefits of the auto-pilot could include not only achieving a better performance but also doing other tasks during the navigation. Further experiment should include a secondary task with the extension of the framework described in subsection 5.1.3 to accommodate the factors that influence the choice behavior, other than performance. The future study should consider the participants' characteristics, such as propensity to trust or attitude

towards robots.

4.4.5 Conclusion

By examining the proposed method with the semi-automatic drone navigation, we demonstrated that our framework with TCCs could promote the trust calibration in the continuous real-time task. The task performance was also improved as a result of the proper trust calibration. Adaptively presenting TCCs was able to change the behaviors of the participants in the situations where the indicator of auto-pilot's reliability failed to maintain the participants' trust for the auto-pilot.

5

Conclusion

5.1 General Discussion

The results of the three empirical studies were obtained with two types of cooperative tasks: discrete and continuous. The overall results indicated that the proposed method could successfully detect and mitigate improper trust calibration.

In this section, we first revisit the applicability of the proposed method in terms of the requirements that possible applications must satisfy. Then, we discuss the Trust Calibration AI (TCAI) regarding its concept and implementation. Finally, we propose an extension to the proposed framework to take care of the trust factors other than performance.

5.1.1 Applicability of Proposed Method

As already described in section 3.2, the applications of the proposed method must meet the following conditions: interchangeability, repeated selection problems, and a

performance-centric view of trust.

(1) Interchangeability

The human user and the AI system should be functionally interchangeable in performing cooperative tasks. Although this requirement may sound too restrictive, there are many real applications for this form of cooperation, including autonomous driving (SAE Level 4) and diagnosis assistance systems. The performance of human users and AI systems varies depending on the tasks and environmental conditions. Despite recent progress on AI technology that exceeds human performance in some tasks [120], humans still outperform AI in many areas that require generalization [114], creativity, and ambiguity. If humans and AI cooperate by understanding the relative strengths and weaknesses of each, the performance of human-AI cooperation would be better than what they could achieve by themselves. The proposed method does not cover classes of applications such as fully autonomous driving (SAE level 5) or AI doctors replacing human doctors. In these applications, the target tasks are performed by AI agents only, and what human users can do is accept or reject the results of the tasks done by the AI agents.

(2) Repeated Selection Problems

Human-AI cooperation should be executed as a series of actions taken by a human user and AI system repeatedly working on selection problems to decide on either AI execution or manual execution. Applications such as visual inspection, cooperative decision making, and cooperative vehicle navigation can be naturally decomposed into such repeated selection problems. This conceptualization is essentially equivalent to a dynamic two-armed bandit problem, which has a wide variety of applications involving making a choice between two alternatives. The system transparency interface of an AI system, which discloses information on AI systems, can help human users solve selection problems. Although both a human user and an AI system can execute a cooperative task, the selections are always decided by the human user who must take final responsibility for the outcome of the cooperation.

(3) Performance-centric View of Trust

As described in Chapter 2, there are many factors influencing trust. In the proposed framework, 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.

There are two types of performance for a human user and AI system to estimate: their performance and the partner's performance. Regarding the self-estimation of one's performance, previous research [121] indicated that human users could appropriately judge their own manual performance, which corresponds to P_H . The TCAI knows how the Task-AI works, so it could use the system information to calculate the Task-AI performance, P_A . Regarding a partner's performance, \widehat{P}_H could be estimated by the TCAI with a model-based or statistical approach using data collected beforehand or on-the-fly during the cooperation. However, the estimation of \widehat{P}_A by a human user could be more complicated even if the trust is decided only on the basis of performance-related factors because humans are known to be prone to cognitive biases. If the factors other than performance-related ones are influencing trust decisions, the extended framework with utility functions described in section 5.1.3 would be required to solve such problems.

As a summary of the discussion of the three requirements, Table 5.1 shows the typical applications of human-AI cooperation. The first three applications satisfy the three requirements described above. In the last two applications, humans and AI have different roles, and they are not functionally interchangeable; therefore, these applications are beyond the scope of the proposed method.

5.1.2 Role of Trust Calibration AI

As described in Chapter 3, the Trust Calibration AI (TCAI) is a conceptual entity that monitors the human user's choice behaviors and issues a trust calibration cue (TCC) if the result estimated by the TCAI does not match the user's selection decision.

To illustrate the characteristics of the TCAI, we discuss comparisons with the cooperation frameworks proposed in previous studies on human-agent teaming.

Table 5.1: Typical applications and their compliance with requirements

Typical Applications	Human's Role	Requirements		
		(1) Interchangeability	(2) Selection problems	(3) Performance-centric view of trust
Security Inspections	Inspector	Yes	Auto or Manual	Yes
Cooperative Medical Diagnosis	Doctor	Yes	Accept or Manual	Yes
Autonomous Driving (Lv4)	Driver	Yes	Auto or Manual	Yes
AI Doctor	Patient	No (Task done by AI only)	Accept or Reject	N/A
Autonomous Driving (Lv5)	Passenger	No (Task done by AI only)	Accept or Stop	N/A

Vecht et al. [122, 123] 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 method with the TCAI.

Cummings et al. [124] 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 method, 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. Note that the decider in our method is always the human user.

The TCAI is designed to not have any identity observable by human users. Thus, there is no trust issue with the TCAI itself because users are not aware that it exists.

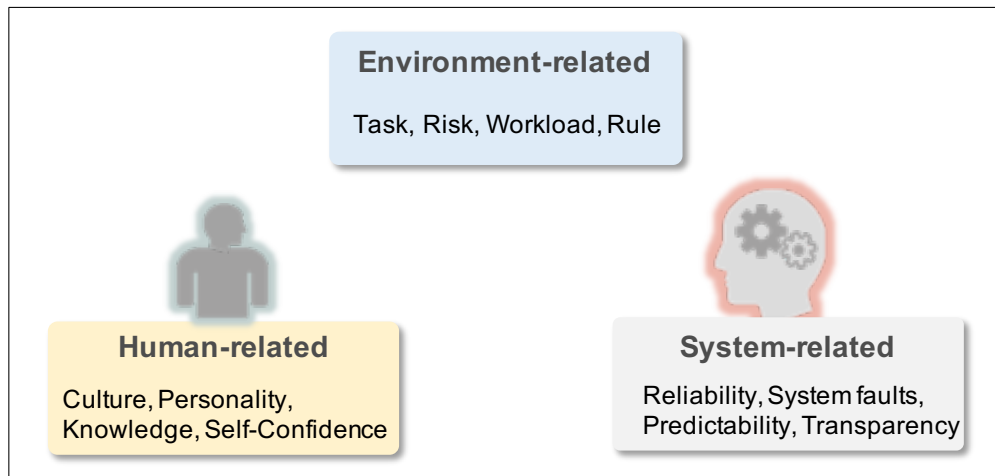


Figure 5.1: Three categories of factors influencing trust decisions

5.1.3 Extension to Framework

We discuss an extension to the proposed framework with expected utility functions to incorporate trust factors other than performance-related ones.

Other Factors Influencing Trust Decisions

The proposed framework focuses on performance-related factors to detect over-trust and under-trust. The reliability of human users and AI systems is compared to identify better selection that might lead to higher performance. However, there are other factors influencing trust decisions in human-AI cooperation, which would make the resulting behaviors for selection problems difficult to understand within the scope of the proposed method defined in Chapter 3. As described in Chapter 2, there are three major categories of trust influencing factors: system-related, human-related, and environment-related. Factors related to performance can be found in the system-related category and also in the human-related category (see Figure 5.1).

Trust decisions could be influenced by the factors classified in the human-related category such as culture, personal traits, and prior knowledge about the AI systems. For example, if a human user has a strong tendency to trust in AI or “automation bias” [125], the selection decision would be inclined toward the use of the Task AI, even when its performance is lower than that of the user.

Task-related factors are included in the environment-related category. For example, one of a driver's motivations for selecting autonomous driving mode could be a possible engagement in non-driving related tasks. Naujoks et al. [126] discussed the desire to engage in non-driving-related tasks during autonomous driving, which requires the workload of the driving task to be lighter and causes the driver to select autonomous mode. Large et al. [127] reported that participants in experiments with a simulated autonomous driving carried out a wide variety of non-driving related tasks such as reading books, engaging in social networking, watching movies, or browsing the web. Even if the reliability of autonomous driving is not higher, a driver may want to engage in a non-driving task and trust autonomous driving just enough to succeed.

Extension with Utility Function

To integrate the other decision factors described above, we consider an extension to the proposed framework based on utility theory [128]. Suppose we have a selection problem S as a set of selections $s_{1:n}$. If $r_{1:n}$ is a set of outcomes and $p_{1:n}$ are their associated probabilities, then a selection s in the selection problem S is written as $s = [r_1 : p_1; \dots; r_n : p_n]$, and an expected utility is given by the following.

$$EU(s) = \sum_{i=1}^n U(r_i) \cdot p_i$$

A utility function $U(x)$ is a real-valued function that represents the degree of desirability of different outcomes r_i .

The selection problem $S_{AI-human}$ in the proposed method described in section 3.2 has two selections: $S_{AI-human} = \{s_{AI}, s_{human}\}$. Suppose that we have n outcomes $a_{1:n}$ and $h_{1:n}$ in s_{AI} and s_{human} , respectively.

$$s_{AI} = [a_1 : p_{a_1}; \dots; a_n : p_{a_n}] \quad (5.1)$$

$$s_{human} = [h_1 : p_{h_1}; \dots; a_n : p_{h_n}] \quad (5.2)$$

We can re-define the second inequalities in the proposed framework (3.1) and (3.2) by

replacing P_A and P_H with $EU(s_{AI})$ and $EU(s_{human})$, respectively.

$$EU(s_{AI}) > EU(s_{human}) \quad (5.3)$$

$$EU(s_{AI}) < EU(s_{human}) \quad (5.4)$$

, where $EU(s_{AI}) = \sum_{i=1}^n U(a_i) \cdot p_{a_i}$ and $EU(s_{human}) = \sum_{i=1}^n U(h_i) \cdot p_{h_i}$.

The outcomes can be defined by whether the task done by AI or human is done successfully or not. Let a_1 be a successful outcome and a_2 be a failed one due to the AI. Similarly, let h_1 and h_2 be a successful outcome and a failed one due to human users, respectively. Now, s_{AI} and s_{human} are

$$s_{AI} = [a_1 : p_{a_1}; a_2 : p_{a_2}] = [a_1 : P_A; a_2 : (1 - P_A)] \quad (5.5)$$

$$s_{human} = [h_1 : p_{h_1}; h_2 : p_{h_2}] = [h_1 : P_H; h_2 : (1 - P_H)] \quad (5.6)$$

, where P_A and P_H are the same as defined in section 3.2. The second inequalities of the proposed framework (3.1) and (3.2) are given as special cases of (5.3) and (5.4) in which $n = 2$ and $U(a_1) = U(h_1) = 1$ and $U(a_2) = U(h_2) = 0$.

Although further research and empirical evaluations should be done to justify this extension concept, it would enable the framework to take into account other factors than performance in human-AI cooperation.

Application Examples

With the extended version of the proposed method (5.3) and (5.4), we show three application examples in the area of autonomous driving (SAE Lv4). The first one is in the human-related category, and the other two are in the environment-related category.

Fun to drive (personal preference): Some drivers show a marked preference for manual driving. A higher value of utility functions $U(h_i)$ can express the situation in which drivers select manual driving mode even if the reliability of autonomous driving P_A is higher than that of manual driving P_H .

No autonomous driving zone (traffic regulation): Traffic situations such as road work, geofences, or accidents are often defined as no-autonomous-driving zones, which prevents autonomous vehicles from staying in autonomous driving mode. Drivers must follow the signal sent from a traffic control system to change from

autonomous to manual driving. The utility functions $U(a_i)$ for the selections of the AI should become zero if the vehicle enters a no-autonomous-driving zone.

Non-driving related task (secondary task): Suppose a driver of an autonomous vehicle would not manually drive because he or she wants to do a secondary task (denoted as t) such as texting on a mobile phone. A new selection s'_{AI} is necessary to represent a situation in which the AI does the driving and the driver does texting. Let t_1 be a successful outcome, and t_2 be a failed one of the secondary task execution by the driver. s'_{AI} has four possible outcomes depending on the successes of the AI task and the secondary task by the driver.

$$s'_{AI} = [a'_1 : p_{a'_1}; a'_2 : p_{a'_2}; a'_3 : p_{a'_3}; a'_4 : p_{a'_4}] \quad (5.7)$$

$$= [a_1, t_1 : p_{a_1 t_1}; a_1, t_2 : p_{a_1 t_2}; a_2, t_1 : p_{a_2 t_1}; a_2, t_2 : p_{a_2 t_2}]; \quad (5.8)$$

The expected utility is $EU(s'_{AI}) = \sum_{i=1}^4 U(a'_i) \cdot p_{a'_i}$. The expected utility is $EU(s'_{AI}) = \sum_{i=1}^4 U(a'_i) \cdot p_{a'_i}$. Now, we can define a new selection problem, $S'_{AI-human} = \{s_{AI}, s'_{AI}, s_{human}\}$, and the rational selection would be the selection with the highest value in terms of the expected utility among $EU(s_{AI})$, $EU(s'_{AI})$, and $EU(s_{human})$.

5.2 Conclusion

This dissertation focused on the problem of over-trust and under-trust in human-AI cooperation by exploring two research questions: RQ1: Can we detect if a user is over-trusting or under-trusting an AI system? RQ2: Can we mitigate a user's over-trust or under-trust? To address these questions, we first examined the related trust literature with a particular focus on trust calibration and factors influencing trust. Measuring trust is essential to detecting miscalibration. To mitigate over-trust or under-trust, influencing trust is necessary. Both measuring trust and influencing trust are difficult because trust is a latent and multi-faceted construct.

We approached the research challenges with a behavior-based trust measurement to capture the status of calibration and a concept of cognitive cues called "trust calibration cues." A formal framework is proposed so that the status of miscalibration can be defined and detected through the observation of human behavior. Four types of trust calibration cues were designed and evaluated. Three empirical studies were

done to evaluate the proposed method. We created two sample tasks for human-AI cooperation: an image screening task and a continuous cooperative navigation task. We conducted three online experiments using a simulated drone environment. We observed both the status of over-trust and the under-trust for the participants of all three experiments. The results of the first empirical study demonstrated that our proposed method had significant effects on changing human behavior in the case of over-trust. A verbal cue showed the largest effect amongst the other cues of visual, audio, and anthropomorphic. The second empirical study showed that the proposed method also worked well under dynamic trust changes of ABA and BAB, where A and B mean over-trust and under-trust. For the image screening task, the level of over-trust was higher than that of under-trust. The third empirical study indicated that the proposed method was effective in a continuous real-time task involving navigating a semi-autonomous drone. This result can open the possibility of applying the proposed method to practical real-time applications such as autonomous driving. We also discussed a possible extension to the framework with utility functions to incorporate trust factors other than performance.

The recent proposal of Trust Engineering for human-AI teaming by Ezer et al. [129] 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. This dissertation's results contribute to the first three components, which are mainly related to interactions between humans and AI.

The results of our empirical evaluations indicated that the proposed method could detect and mitigate the status of improper trust calibration; therefore, we conclude that our proposed method provides a reasonable basis for answering the two research questions, RQ1 and RQ2. As the proposed method is based on a simple and task-independent framework, it could be applied to many application situations. Despite several limitations, we believe that our proposed method could contribute to a baseline design of trustworthy systems for better human-AI cooperation.

5.3 Future Work

As described in Chapter 2, Lee and See [1] proposed three categories of factors influencing trust: performance, purpose, and process. Hoffman et al. [31, 9] also studied three categories: human-related, robot-related, and environment-related. This dissertation discussed trust calibration with a particular focus on performance aspects of human-AI cooperation. There are many other factors influencing trust to be considered in future research. Human characteristics such as age, gender, and propensity to trust should be examined. Trust could also be significantly impacted by attributes of AI systems or autonomous robots such as appearance and anthropomorphism. Further evaluation of the proposed method with different types of robots and tasks should be conducted.

Although we have learned some lessons in the empirical study indicating that the TCCs were more effective than a simple reliability indicator in the case of miscalibration, further research on the interactions between human users and AI systems is required to evaluate the concept of TCCs. Regarding the TCAI, which is a conceptual entity in our proposed method, implementing an explicit interface/appearance for the TCAI would be interesting to explore. Although this might bring about new issues such as regarding trust in the TCAI, it would help us to acquire a better understanding of the trust calibration mechanism in human-AI cooperation.

Bibliography

- [1] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.
- [2] Kevin Hoff and Masooda Bashir. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3):407–434, 2015.
- [3] Kim Drnec, Amar R Marathe, Jamie R Lukos, Jason S Metcalfe, Klaus Gramann, Agnieszka Wykowska, and Dietrich Manzey. From Trust in Automation to Decision Neuroscience: Applying Cognitive Neuroscience Methods to Understand and Improve Interaction Decisions Involved in Human Automation Interaction. *Frontiers In Human Neuroscience*, 10(290), 2016.
- [4] Jason M. Bindewald, Christina F. Rusnock, and Michael E. Miller. Measuring Human Trust Behavior in Human-Machine Teams. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, volume 591, pages 47–58, 2018.
- [5] Tove Helldin, Göran Falkman, Maria Riveiro, and Staffan Davidsson. 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*, pages 210–217, 2013.
- [6] Renate Haeuslschmid, Max Von Buelow, B. Pfleging, and A. Butz. Supporting Trust in Autonomous Driving. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 319–329, 2017.

- [7] Denise M. Rousseau, Sim B. Sitkin, Ronald S. Burt, and Colin Camerer. Introduction to Special Topic Forum : Not so Different after All : A Cross-Discipline View of Trust Denise M . Rousseau ; Sim B . Sitkin ; Ronald S . Burt ; Colin Camerer. *The Academy of Management Review*, 23(3):393–404, 1998.
- [8] Bonnie M. Muir. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922, 1994.
- [9] Kristin E. Schaefer, Jessie Y.C. Chen, James 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. *Human Factors*, 58(3):377–400, 2016.
- [10] Stephan Lewandowsky. The dynamics of trust : Comparing humans to automation. *Experimental Psychology Applied*, 6(2):104–123, 2000.
- [11] Bonnie M. Muir and Neville Moray. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460, 1996.
- [12] National Transportation Safety Board (NTSB). Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator. Technical report, 2018.
- [13] National Transportation Safety Board (NTSB). Preliminary Report - Crash Involving Pedestrian - Uber Test Vehicle. Technical report, 2018.
- [14] National Highway Traffic Safety Administration (NHTSA). Automatic vehicle control systems – investigation of Tesla accident. Technical report, 2017.
- [15] Ewart J. de Visser, Marieke M.M. Peeters, F. Jung Malte, Spencer Kohn, H. Shaw Tyler, Richard Pak, and Mark A Neerincx. Towards a Theory of Longitudinal Trust Calibration in Human-Robot Teams. *International Journal of Social Robotics*, pages 1–20, 2019.

- [16] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of Robot Failures and Feedback on Real-Time Trust. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 251–258, 2013.
- [17] Kristin E. Oleson, D. R. Billings, Vivien Kocsis, Jessie Y.C. Chen, and P. A. Hancock. Antecedents of trust in human-robot collaborations. In *Proceedings of the 2011 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support*, number March, pages 175–178, 2011.
- [18] Bonnie M Muir. Trust between humans and machines. *International Journal of Man-Machine Studies*, 27:527–539, 1987.
- [19] John K. Rempel, John G. Holmes, and Mark P. Zanna. Trust in Close Relationships. *Journal of Personality and Social Psychology*, 49(1):95–112, 1985.
- [20] John M. McGuirl, Nadine B. Sarter, John M. McGuirl, and Nadine B. Sarter. Supporting Trust Calibration and the Effective Use of Decision Aids by Presenting Dynamic System Confidence Information. *Human Factors*, 48(4):656–665, dec 2006.
- [21] Ewart J de Visser, Marvin Cohen, Amos Freedy, and Raja Parasuraman. A Design Methodology for Trust Cue Calibration in Cognitive Agents. In *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality*, pages 251–262, 2014.
- [22] Tove Helldin. *Transparency for Future Semi-Automated Systems*. Doctoral dissertation, Orebro University, 2014.
- [23] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. Taxonomy of trust-Relevant Failures and Mitigation Strategies. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 3–12, 2020.
- [24] James Llinas, Ann M Bisantz, Colin G Drury, Y. Seong, Jiun-Yin Y. Jian, Y. Seong, and Jiun-Yin Y. Jian. Studies and Analyses of Aided Adversarial Decision Making.

- Phase 2: Research on Human Trust in Automation. Technical report, Air Force Research Laboratory, 1998.
- [25] Roger C Mayer, James H Davis, F David Schoorman, Roger C Mayer, and James H Davis. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709–734, 1995.
- [26] Rajeev Bhattacharya, Timothy M. Devinney, and Madan M. Pillutla. A Formal Model of Trust Based on Outcomes. *The Academy of Management Review*, 23(3):459–472, 1998.
- [27] Maria Madsen. Measuring Human-Computer Trust. In *Proceedings of the 11th Australasian Conference on Information Systems*, pages 6–8, 2000.
- [28] Michael Lewis, Katia Sycara, and Phillip Walker. The Role of Trust in Human-Robot Interaction. *Studies in Systems, Decision and Control*, 117:135–159, 2018.
- [29] Barbara D Adams, Lora E Bruyn, and Sébastien Houde. Trust in Automated Systems Literature Review. Technical report, Defence Research and Development Canada, Toronto, Canada, 2003.
- [30] Bronwyn French, Andreas Duenser, and Andrew Heathcote. Trust in Automation: A literature review. Technical report, CSIRO, Australia., 2018.
- [31] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53(5):517–527, 2011.
- [32] Dingjun Li, P L Patrick Rau, and Ye Li. A Cross-cultural Study : Effect of Robot Appearance and Task. *International Journal of Social Robotics*, 2(April):175–186, 2010.
- [33] Shih-yi Chien, Michael Lewis, Katia Sycara, and Jyishane Liu. The Effect of Culture on Trust in Automation : Reliability and Workload. *ACM Transactions on Interactive Intelligent Systems*, 8(4):1–31, 2018.

- [34] Geoffrey Ho, Dana Wheatley, and Charles T Scialfa. Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6):690–710, 2005.
- [35] Frederick Steinke, Tobias Fritsch, Daniel Brem, and Svenja Simonsen. Requirement of AAL systems – Older persons’ trust in sensors and characteristics of AAL technologies. In *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive*, pages 1–6, 2012.
- [36] James L Szalma and Grant S Taylor. Individual Differences in Response to Automation : The Five Factor Model of Personality. *Experimental Psychology Applied*, 17(2):71096, 2011.
- [37] Stephanie M Merritt, Missouri St, and Daniel R Ilgen. Not All Trust Is Created Equal : Dispositional and History- Based Trust in Human-Automation Interactions. *Human Factors*, 50(2):194–210, 2008.
- [38] Nora Balfe, Trinity College Dublin, Sarah Sharples, and John R Wilson. Understanding Is Key : An Analysis of Factors Pertaining to Trust in a Real-World Automation System. *Human Factors*, 60(4):477–495, 2018.
- [39] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, 15(2):134–160, 2014.
- [40] John D. Lee and Neville Moray. Trust, self-Confidence, and operators’ adaptation to automation. *International Journal of Human-Computer Studies*, 40(1):153–184, 1994.
- [41] Peter de Vries, Cees Midden, and Don Bouwhuis. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, 58(6):719–735, 2003.
- [42] Xiaocong Fan, Michael Mcneese, John Yen, and Haydee Cuevas. The Influence of Agent Reliability on Trust in Human-Agent Collaboration. In *Proceedings of the 15th European conference on Cognitive ergonomics*, pages 1–8, 2008.

- [43] Alyssa Glass, Deborah L Mcguinness, and Michael Wolverton. Toward Establishing Trust in Adaptive Agents. In *Proceeding of the 13th International Conference on Intelligent User Interfaces*, pages 227–236, 2008.
- [44] Stuart Moran, Nadia Pantidi, Khaled Bachour, Joel E Fischer, Martin Flintham, Tom Rodden, Simon Evans, and Simon Johnson. Team Reactions to Voiced Agent Instructions in a Pervasive Game. In *Proceedings of the 18th International Conference on Intelligent User Interfaces*, pages 371–382, 2013.
- [45] Randall D. Spain and James P. Bliss. The effect of sonification display pulse rate and reliability on operator trust and perceived workload during a simulated patient monitoring task. *Ergonomics*, 51(9):1320–1337, 2008.
- [46] Neville Moray, Inagaki Toshiyuki, and Itoh Makoto. Adaptive Automation , Trust , And Self-Confidence in Fault Management of Time-Critical Tasks. *Experimental Psychology: Applied*, 6(1):44–58, 2000.
- [47] John Lee and Neville Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–1270, 1992.
- [48] Kun Yu, Ronnie Taib, and Fang Chen. User Trust Dynamics : An Investigation Driven by Differences in System Performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317, 2017.
- [49] Robert R. Hoffman, Gary Klein, and Shane T. Mueller. EXPLAINING EXPLANATION FOR "EXPLAINABLE AI". In *Proceedings of the Human Factors and Ergonomics Society*, volume 1, pages 197–201, 2018.
- [50] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. Evaluating Effects of User Experience and System Transparency on Trust in Automation. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 408–416, 2017.
- [51] Leeann Perkins, Janet E Miller, Ali Hashemi, and Gary Burns. Designing for Human-Centered Systems : Situational Risk as a Factor of Trust in Automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 2130–2134, 2010.

- [52] Alan R. Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems*, 8(4), 2018.
- [53] David P. Biro, Mark Daly, and Gregg Gunsch. The Influence of Task Load and Automation Trust on Deception Detection. *Group Decision and Negotiation*, 13(2):173–189, 2004.
- [54] David L Strayer and William A Johnston. DRIVEN TO DISTRACTION: Dual-Task Studies of Simulated Driving and Conversing on a Cellular Telephone. *Psychological Science*, 12(6):462–466, 2001.
- [55] Frederik Naujoks, Dennis Befelein, Katharina Wiedemann, and Alexandra Neukum. A Review of Non-driving-related Tasks Used in Studies on Automated Driving. *Advances in Human Aspects of Transportation*, 1:525–537, 2018.
- [56] Jason Metcalfe, Amar Marathe, B Haynes, V J Paul, G M Gremillion, K Drnec, C Atwater, J R Estepp, J R Lukos, E C Carter, and W D Nothwang. Building a framework to manage trust in automation. *Micro-and Nanotechnology Sensors, Systems, and Applications IX*, 10194, 2017.
- [57] Peter A Hancock. Can You Trust Your Robot? *Ergonomics in Design*, 19(3):24–29, 2011.
- [58] Ji Gac and John D. Lee. Extending the Decision Field Theory to Model Operators' Reliance on Automation in Supervisory Control Situations. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, 36(5):943–959, 2006.
- [59] Jiun-yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71, 2000.
- [60] P. Goillau, C. Kelly, M. Boardman, and E. Jeannot. Guidelines for Trust in Future ATM Systems : Measures. Technical report, the European Organization for the Safety of Air Navigation, 2003.

- [61] Rosemarie E Yagoda and Douglas J Gillan. You Want Me to Trust a ROBOT? The Development of a Human–Robot Interaction Trust Scale. *International Journal of Social Robotics*, 4(3):235–248, 2012.
- [62] David Miller, Mishel Johns, Brian Mok, Nikhil Gowda, David Sirkin, Key Lee Twitter, and Wendy Ju. Behavioral Measurement of Trust in Automation: The Trust Fall. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 1849–1853, 2016.
- [63] William Payre, Julien Cestac, and Patricia Delhomme. Fully Automated Driving: Impact of Trust and Practice on Manual Control Recovery. *Human Factors*, 58(2):229–241, 2016.
- [64] Ji Gao, John D. Leeb, and Yi Zhang. A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics*, 36(5):511–526, 2006.
- [65] Chandrayee Basu and Mukesh Singhal. Trust Dynamics in Human Autonomous Vehicle Interaction : A Review of Trust Models. *2016 AAAI Spring Symposium Series*, pages 85–91, 2016.
- [66] Ewart J. de Visser, Samuel S Monfort, Kimberly Goodyear, Rhode Island, Li Lu, Martin O Hara, Fairfax Hospital, Mary R Lee, and Frank Krueger. A Little Anthropomorphism Goes a Long Way : Effects of Oxytocin on Trust, Compliance, and Team Performance With Automated Agents. *Human Factors*, 59(1):116–133, 2017.
- [67] Sebastian Hergeth, Lutz Lorenz, Roman Vilimek, and Josef F. Krems. Keep Your Scanners Peeled: Gaze Behavior as a Measure of Automation Trust during Highly Automated Driving. *Human Factors*, 58(3):509–519, 2016.
- [68] Parham Nooralishahi, Loo Chu Kiong, Ai-Vyrn Chin, Halimahtun M Khalid, Liew Wei Shiung, Zeeshan Rasool, Martin G Helander, and Chin Ai-vyrn. Exploring Psycho-Physiological Correlates to Trust: Implications for Human-Robot-Human Interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 697–701, 2016.

- [69] Anqi Xu and Gregory Dudek. OPTIMO: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 221–228, mar 2015.
- [70] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *ACM Transactions on Human-Robot Interaction*, 9(2):1–23, 2020.
- [71] Kumar Akash, Katelyn Polson, Tahira Reid, and Neera Jain. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part I: Human Trust and Workload Model. *IFAC-PapersOnLine*, 51(34):315–321, 2019.
- [72] David V. Pynadath, Ning Wang, and Sreekar Kamireddy. A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, pages 171–178, Kyoto, 2019.
- [73] Stephanie M. Merritt. Affective processes in human-automation interactions. *Human Factors*, 53(4):356–370, 2011.
- [74] Andrew C Wicks. The Structure of Optimal Trust : Moral and Strategic Implications. *The Academy of Management Review*, 24(1):99–116, 1999.
- [75] Scott Ososky, David Schuster, Elizabeth Phillips, and Florian Jentsch. Building Appropriate Trust in Human-Robot Teams. *AAAI Spring Symposium Series*, pages 60–65, 2013.
- [76] Raja Parasuraman and Victor Riley. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 2(39):230– 253, 1997.
- [77] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M. Howard, and Alan R. Wagner. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 101–108, 2016.

- [78] Stephanie M. Merritt, Kelli Huber, Jennifer LaChapell-Unnerstall, and Deborah Lee. Continuous Calibration of Trust in Automated Systems. Technical report, Air Force Research Laboratory, 2014.
- [79] Jessie Y.C. Chen and Michael J. Barnes. Human - Agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29, 2014.
- [80] Joseph B. Lyons and Paul R. Havig. Transparency in a Human-Machine Context: Approaches for Fostering Shared Awareness/Intent. In *Proceedings of the International Conference on Virtual, Augmented and Mixed Reality*, pages 181–190, 2014.
- [81] Bobbie D. Seppelt and John D. Lee. Making adaptive cruise control (ACC) limits visible. *International Journal of Human Computer Studies*, 65(3):192–205, 2007.
- [82] Bobbie Danielle Seppelt. *Supporting operator reliance on automation through continuous feedback*. Doctoral dissertation, University of Iowa, 2009.
- [83] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, 58(3):401–415, 2016.
- [84] Joseph B. Lyons, Garrett G. Sadler, Kolina Koltai, Henri Battiste, Nhut T. Ho, Lauren C. Hoffmann, David Smith, Walter Johnson, and Robert Shively. Shaping Trust Through Transparent Design: Theoretical and Experimental Guidelines. *Advances in Human Factors in Robots and Unmanned Systems*, pages 127–136, 2017.
- [85] Kumar Akash, Tahira Reid, and Neera Jain. Improving Human-Machine Collaboration Through Transparency-based Feedback – Part II: Control Design and Synthesis. *IFAC-PapersOnLine*, 2019.
- [86] Jessie Y.C. Chen, Michael J. Barnes, Anthony R. Selkowitz, and Kimberly Stowers. Effects of Agent Transparency on Human-Autonomy Teaming Effectiveness. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1838–1843, 2016.

- [87] Michael W. Boyce, Jessie Y.C. Chen, Anthony R. Selkowitz, and Shan G. Lakhmani. Effects of Agent Transparency on Operator Trust. In *Proceedings of the Tenth ACM/IEEE International Conference on Human-Robot Interaction*, pages 179–180, 2015.
- [88] Kristin E. Schaefer, Ralph W. Brewer, Joe Putney, Edward Mottern, Jeffrey Barghout, and Edward R. Straub. Relinquishing Manual Control - Collaboration Requires the Capability to Understand Robot Intention. In *Proceedings of the 2016 International Conference on Collaboration Technologies and Systems*, pages 359–366, 2016.
- [89] Ning Wang, David V. Pynadath, and Susan G. Hill. Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 109–116, 2016.
- [90] Long Chen, Libo Sun, Teng Yang, Lei Fan, Kai Huang, and Zhe Xuanyuan. RGB-T SLAM: A Flexible SLAM Framework by Combining Appearance and Thermal Information. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation*, pages 5682–5687, 2017.
- [91] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elis-Abeth Elisabeth André. "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, volume 19, pages 7–9, 2019.
- [92] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How Do Visual Explanations Foster End Users' Appropriate Trust in Machine Learning? In *Proceedings of the International Conference on Intelligent User Interfaces*, pages 189–201, 2020.
- [93] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 295–305, 2020.

- [94] Ewart J. de Visser, Richard Pak, and Tyler H. Shaw. From 'automation' to 'autonomy': the importance of trust repair in human-machine interaction. *Ergonomics*, 61(10):1409–1427, 2018.
- [95] Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. Persuasive Robotics: The Influence of Robot Gender on Human Behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2563–2568, 2009.
- [96] Sean Ye, Glen Neville, Mariah Schrum, Matthew Gombolay, Sonia Chernova, and Ayanna Howard. Human Trust After Robot Mistakes: Study of the Effects of Different Forms of Robot Communication. In *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*, pages 1–7, 2019.
- [97] Sandy H. Huang, Kush Bhatia, Pieter Abbeel, and Anca D. Dragan. Establishing Appropriate Trust via Critical States. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 3929–3936, 2018.
- [98] Yuan Gao, Elena Sibirtseva, Ginevra Castellano, and Danica Kragic. Fast Adaptation with Meta-Reinforcement Learning for Trust Modelling in Human-Robot Interaction. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 305–312, 2019.
- [99] Dylan P. Losey, Student Member, and Dorsa Sadigh. Robots that Take Advantage of Human Trust. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 7001–7008, 2019.
- [100] Akihiro Maehigashi, Kazuhisa Miwa, Hitoshi Terai, Kazuaki Kojima, and Junya Morita. Selection Strategy of Effort Control: Allocation of Function to Manual Operator or Automation System. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 1977–1982, 2011.
- [101] David Good. Individuals, Interpersonal Relations, and Trust. In *Trust: Making and Breaking Cooperative Relations*, chapter 3, pages 31–48. University of Oxford, 2000.

- [102] Hua Cai and Yingzi Lin. Tuning Trust Using Cognitive Cues for Better Human-Machine Collaboration. In *Proceedings of the Human Factors and Ergonomics Society*, pages 2437–2441, 2010.
- [103] Takanori Komatsu, Seiji Yamada, Kazuki Kobayashi, Kotaro Funakoshi, and Mikio Nakano. Artificial Subtle Expressions: Intuitive Notification Methodology of Artifacts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1941–1944, 2010.
- [104] Andrew J. Cowell and Kay M. Stanney. Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International Journal of Human Computer Studies*, 62(2):281–306, 2005.
- [105] Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014.
- [106] Kenneth R. Laughery and Michael S. Wogalter. A three-stage model summarizes product warning and environmental sign research. *Safety Science*, 61:3–10, 2014.
- [107] Mark A. Changizi, Matt Brucksch, Ritesh Kotecha, Kyle McDonald, and Kevin Rio. Ecological warnings. *Safety Science*, 61:36–42, 2014.
- [108] Sung Won Lee, SeokJin Kim, Jeong Han, Kwang Eun An, Seung-Ki Ryu, and Dongmahn Seo. Experiment of Image Processing Algorithm for Efficient Pothole Detection. In *Proceedings of the IEEE International Conference on Consumer Electronics*, pages 1–2, 2019.
- [109] Mjc J C Crump, J V McDonnell, and T M Gureckis. Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3):57410, 2013.
- [110] The Cesium Consortium. CesiumJS - Geospatial 3D Mapping and Virtual Globe Platform, 2018.
- [111] Microsoft. Bing Maps API Documentation, 2018.

- [112] Harold Stanislaw. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999.
- [113] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105, 2012.
- [114] Robert Geirhos, Carlos R Medina Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pages 7549–7561, 2018.
- [115] Samuel Dodge and Lina Karam. A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions. In *Proceedings of the 26th International Conference on Computer Communications and Networks*, pages 1–7, 2017.
- [116] Vincent C Conzola and M. S. Wogalter. A Communication – Human Information Processing (C–HIP) approach to warning effectiveness in the workplace. *Journal of Risk Research*, 4(4):309–322, 2001.
- [117] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000.
- [118] Alexander G. Mirnig, Philipp Wintersberger, Christine Sutter, and Jürgen Ziegler. A Framework for Analyzing and Calibrating Trust in Automated Vehicles. In *Adjunct Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 33–38, 2017.
- [119] SAE International. SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. Technical report, 2018.
- [120] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification.

- In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.
- [121] Janet Metcalfe and Matthew Jason Greene. Metacognition of agency. *Journal of Experimental Psychology: General*, 136(2):184–199, 2007.
- [122] 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 *NATO HFM symposium on Human Autonomy Teaming*, pages 1–26, 2018.
- [123] Bob van der Vecht, Jurriaan van Diggelen, Marieke Peeters, Jonathan Barnhoorn, and Jas per van der Waa. SAIL: A Social Artificial Intelligence Layer for Human-Machine Teaming. In *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems*, pages 262–274, 2018.
- [124] Mary L. Cummings and Sylvain Bruni. Collaborative Human–Automation Decision Making. In *Handbook of Automation*, pages 437–447. Springer, 2009.
- [125] Raja Parasuraman and Dietrich H. Manzey. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors*, 52(3):381–410, 2010.
- [126] Frederik Naujoks, Katharina Wiedemann, and Nadja Schömig. The Importance of Interruption Management for Usefulness and Acceptance of Automated Driving. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 254–263, 2017.
- [127] David R. Large, Gary Burnett, Andrew Morris, Arun Muthumani, and Rebecca Matthias. A Longitudinal Simulator Study to Explore Drivers’ Behaviour During Highly-Automated Driving. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, pages 583–594, 2017.
- [128] Mykel J. Kochenderfer. *Decision Making Under Uncertainty: Theory and Application*. MIT Press, 2015.

- [129] Neta Ezer, Sylvain Bruni, Yang Cai, Sam J. Heppenstal, Christopher A. Miller, and Dylan D. Schmorow. Trust Engineering for Human-AI Teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, pages 322–326, 2019.

List of Publication

Journal

1. Kazuo Okamura and Seiji Yamada. Adaptive trust calibration for human-AI collaboration. *PLoS ONE*, 15(2):e0229132, 2020. <https://doi.org/10.1371/journal.pone.0229132>

Conference (peer reviewed)

1. Kazuo Okamura and Seiji Yamada. Adaptive Trust Calibration for Supervised Autonomous Vehicles, In *Adjunct Proceedings of the 10th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 92–97, 2018. <https://doi.org/10.1145/3239092.3265948>
2. Kazuo Okamura and Seiji Yamada. Calibrating Trust in Autonomous Systems in a Dynamic Environment. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 2492-2498, 2020. <https://cognitivesciencesociety.org/cogsci20/papers/0590/0590.pdf>
3. Kazuo Okamura and Seiji Yamada. Calibrating Trust in Human-Drone Cooperative Navigation. In *Proceedings of the 29th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1274-1279, 2020

Conference (not peer reviewed)

1. 岡村和男、山田誠二. 人間とエージェントとの協調作業における適応的な信頼校正手法の提案. 第 27 回人工知能学会全国大会,1G4-OS-13b-03, 4

pages, 2019. https://doi.org/10.11517/pjsai.JSAI2019.0_1G4OS13b03