PAPER

Experimental Investigation of Calibration and Resolution in Human-Automation System Interaction

Akihiro MAEHIGASHI[†a)], ***Nonmember***, **Kazuhisa MIWA**[†], ***Member***, **Hitoshi TERAI**[††], **Kazuaki KOJIMA**[†††*], ***Nonmembers***, ***and*** **Junya MORITA**[††††], ***Member***

SUMMARY This study investigated the relationship between human use of automation and their sensitivity to changes in automation and manual performance. In the real world, automation and manual performance change dynamically with changes in the environment. However, a few studies investigated whether changes in automation or manual performance have more effect on whether users choose to use automation. We used two types of experimental tracking tasks in which the participants had to select whether to use automation or conduct manual operation while monitoring the variable performance of automation and manual operation. As a result, we found that there is a mutual relationship between human use of automation and their sensitivity to automation and manual performance changes. Also, users do not react equally to both automation and manual performance changes although they use automation adequately.

key words: *human-automation system interaction, Misuse, Disuse, calibration, resolution*

1. Introduction

1.1 Human Use of Automation

In human-automation system interaction, users perform supervisory control [1]–[6]. The primary role of supervisory control is to determine whether to let the automated system perform the task or to manually perform it; this is done by monitoring the performance of automation and manual operation. However, Parasuraman and Riley [7] showed that misselections in supervisory control have occurred and caused fatal accidents. They defined such misselections as misuse (overreliance on automation) and disuse (underreliance on automation).

Many studies have experimentally investigated human use of automation. Some showed that people tend to misuse automation [8]–[11]. In these studies, multiple-task situations were set up for the experiments, and participants had to perform multiple tasks manually except for one automated task. The results indicated that in such a situation, participants tended to allocate their attentional resources to the manual tasks and not to monitoring of the automation performance. Because of this neglect of automation performance monitoring, called complacency, participants could not recognize the performance errors of the automated system and fell into automation misuse [12]. In other words, these studies showed that users tend to misuse automation because of their insensitivity to degradation in automation performance.

Conversely, other studies showed that people tend to disuse automation [13]–[16]. In these studies, diagnostic tasks were used as experimental tasks. Participants had to select whether to let automated decision support systems perform the tasks or to manually perform the tasks by themselves by comparing the performance of the system and of manual operation. These studies indicated that in such a situation, even when the automated support systems showed nearly perfect task performance, the participants reacted sensitively to system errors and fell into automation disuse. This preconception of automation as a perfect aid is called the perfect automation schema [15]. Thus, these studies showed that users tend to disuse automation because they are too sensitive to degradations in its performance.

In the real world, automation and manual performance change dynamically with changes in the environment [7]. However, a few studies investigated whether users react more sensitively to performance changes in automation or in manual operation when they decide whether to use automation. They might make the decision while reacting more sensitively to changes in one than in the other or while reacting equally to changes in both. In this study, we experimentally investigate the relation between human use of automation, the tendency to fall into misuse or disuse, and sensitivity to changes in the performance of automation and manual operation.

1.2 Calibration and Resolution

We measured and evaluated participants' use of automation and their sensitivity to performance changes using the concepts of calibration and resolution. "Calibration refers to the correspondence between a person's trust in the automation and the automation's capabilities. [partially omitted] Resolution refers to how precisely a judgment of trust differentiates levels of automation capability" [17, pp.55, 56]. In other words, calibration indicates the relationship between a

Manuscript received January 11, 2013.
Manuscript revised February 26, 2013.

[†]The authors are with the Graduate School of Information Science, Nagoya University, Nagoya-shi, 464-8601 Japan.
[††]The author is with JST-CREST, Tokyo, 102-0075 Japan.
[†††]The author was with the Faculty of Human Sciences, Waseda University, Tokorozawa-shi, 359-1192 Japan.
[††††]The author is with School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi-shi, 923-1292 Japan.
[*]Presently, with Learning Technology Laboratory, Teikyo University, Utsunomiya-shi, 320-8551 Japan.
a) E-mail: mhigashi@cog.human.nagoya-u.ac.jp

DOI: 10.1587/transfun.E96.A.1625

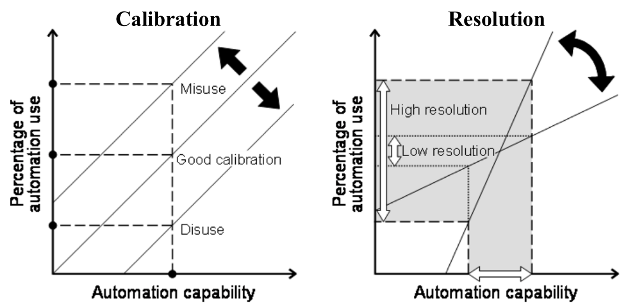

Fig. 1 Conceptualized images of calibration and resolution. The x-axis represents automation capability, and the y-axis represents percentage of automation use.

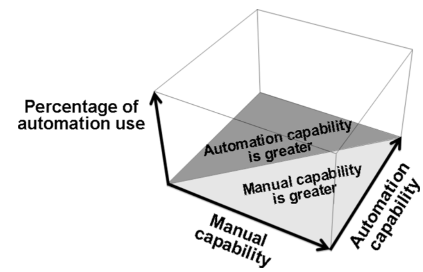

Fig. 2 3D graph showing relationship between automation and manual capabilities. The x-axis represents the manual capability as a dependent variable, the y-axis represents the automation capability as a dependent variable, and the z-axis represents the percentage of automation use as an independent variable.

user's trust in automation and a particular automation capability. "Overtrust is poor calibration in which trust exceeds system capabilities; with distrust, trust falls short of the automation's capabilities" [17, pp.55]. Conversely, resolution defines the relationship between changes in a user's trust in automation and automation capability. "Poor resolution occurs when a large range of automation capability maps onto a small range of trust. With low resolution, large changes in automation capability are reflected by small changes in trust" [17, pp.56].

In this study, we discuss human behavior on the basis of human use of automation rather than trust-based human inner states in order to investigate human reaction to automation in a changing environment. We conceptualize the relationship between human use of automation and automation capability using calibration and resolution. Figure 1 shows conceptualized images of calibration and resolution. By using these graphs, we can measure and evaluate users' tendency to use automation and their sensitivity to changes in automation performance. Calibration indicates the tendency to use automation. Good calibration shows a correspondence between human use of automation and automation capability. In misuse, the line on the graph shifts upward, indicating overreliance on automation; in disuse, the line shifts downward, showing underreliance on automation. Moreover, resolution indicates the sensitivity to changes in automation performance. At high resolution, the change in automation use is steeper with changes in automation capability, showing a sensitive reaction to changes in automation performance. Conversely, at low resolution, the change in automation use is more gradual with changes in automation capability, showing an insensitive reaction to changes in automation performance.

However, it is important to consider the relationship between the performance of automated systems and manual operation to evaluate users' tendency to use automation [2]–[4]. The efficiency of automation use differs depending on this relationship. When automation outperforms manual operation, it is preferable to use automation. In contrast, when manual operation outperforms automation, it is preferable to conduct manual operation. Therefore, in this study, we evaluate users' automation usage on the basis on the relationship between automation and manual performance using the concept of calibration. In addition, we consider users' sensitivity not only to changes in automation performance but also to changes in manual performance using the concept of resolution.

The 3D graph in Fig. 2 shows the capabilities of automation and manual operation as dependent variables and the percentage of automation use as an independent variable. We assume that when the automation and manual capabilities are equal, automation and manual operation exhibit identical performance. Figure 2 shows the difference in the efficiency of using automation and conducting manual operations depending on the relationship between the automation and manual capabilities. The efficiency is different on either side of the boundary line where the automation and manual capabilities are equal. In front of the line (light gray area), the manual capability is greater than the automation capability, and it is efficient to conduct manual operation. In contrast, behind the line (dark gray area), the automation capability is greater than the manual capability, and it is efficient to use automation. In this study, we evaluate users' calibration and resolution on the basis of the percentage of automation use, as shown in Fig. 2.

1.3 Calibration and Resolution in This Study

In this study, we define calibration as the adjustment in automation use based on the relationship between the automation and manual capabilities. Figure 3 shows graphical images of three types of calibration. We define good calibration as adequate use of automation, i.e., neither misuse nor disuse. In addition, we define misuse-biased calibration as the tendency toward overreliance on automation; in contrast, we define disuse-biased calibration as the tendency toward underreliance on automation and overreliance on manual operation. As shown in Fig. 3, calibration is evaluated on the basis of the vertical position of the graphical surface, which shows the percentage of automation use. When calibration is good, automation tends to be used in situations where it is efficient, and manual operation tends to be conducted in situations where it is efficient. In this case, the percentage of automation use, depicted as the curved surface, shows no bias in the vertical position. However, in misuse bias, the

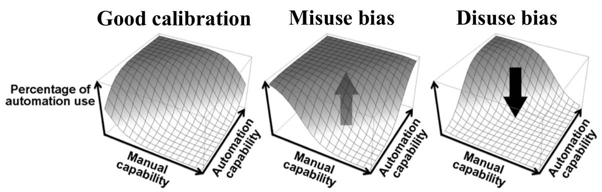

Fig. 3 Graphical images of three types of calibration. In each image, the x-axis represents the manual capability, the y-axis represents the automation capability, and the z-axis represents the percentage of automation use. We assume that the rate of increase of manual and automation capabilities on the x and y axes, respectively, are identical.

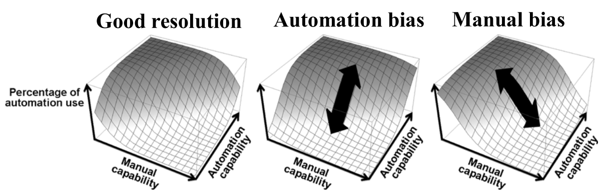

Fig. 4 Graphical images of three types of resolution. In each image, the x-axis represents the manual capability, the y-axis represents the automation capability, and the z-axis represents the percentage of automation use. We assume that the rate of increase of manual and automation capabilities on the x and y axes, respectively, are identical.

graphical surface shifts upward, showing a tendency to use automation even when manual operation is efficient. In disuse bias, the graphical surface shifts downward, showing a tendency to conduct manual operation even when automation use is efficient.

We define resolution as users' sensitivity to changes in the automation and manual capabilities. Figure 4 shows graphical images of three types of resolution. In this study, we investigate whether users react more sensitively to changes in the performance of automation or manual operation when they decide whether to use automation. We define unbiased resolution as equal user sensitivity to changes in both. In addition, we define automation-biased resolution as greater user sensitivity to changes in automation performance than to changes in manual performance. Conversely, we define manual-biased resolution as greater user sensitivity to changes in manual performance than to changes in automation performance. As shown in Fig. 4, the resolution is evaluated on the basis of the slope angle of the graphical surface, which shows the change in the percentage of automation use. In unbiased resolution, the graphical surface is horizontally flat; that is, the percentage of automation use changes equally with changes in the automation and manual capabilities. It means that users react equally to these changes when they decide whether to use automation or conduct manual operation. However, in automation bias, the slope angle of the graphical surface is steeper for changes in automation capability, showing that users react more sensitively to these changes than to changes in manual capability when choosing whether to use automation. Conversely, in manual bias, the slope angle of the graphical surface is steeper for changes in manual capability, showing that users react more sensitively to these changes than to changes in automation capability when choosing whether to use automation. Theoretically, calibration and resolution as defined in this study are considered to be independent. However, a mutual dependence mediated by human psychological factors might exist between calibration and resolution.

Some studies experimentally investigated human use of automation based on the performance of both automation and manual operation using diagnostic tasks [13]–[16]. In these studies, an automated decision support system was used. The automated system would make twice or half as many correct decisions as made by the participants. The participants had to evaluate the performances of the automated system and their manual operation, and they had to choose whether to use the automated system or conduct the task manually. However, these studies did not consider dynamic changes in both automation and manual performance. In addition, Madhavan and Phillips [18] experimentally investigated human sensitivity to automation performance changes using a diagnostic task. They indicated that there are individual differences in the ability to change the selection of automation use with automation performance changes. However, they experimentally manipulated only changes in the automation performance and did not consider changes in the manual performance. Moreover, de Vries, Midden, and Bouwhuis [19] experimentally investigated human use of automation using the concepts of calibration and resolution defined in this study by employing a route planning task. They manipulated two factors, the automation and manual capabilities, with two performance levels each (higher- and lower-performance groups). They found that participants tended to disuse automation and decided whether to use automation while reacting more sensitively to changes in automation performance than to changes in manual performance. However, because they manipulated two factors in between participants design, their participants did not actually experience the performance changes during the task.

In contrast to the previous studies, in this study we set up a situation where participants must monitor both the automation and manual performance as they change dynamically and select whether to use automation or conduct manual operation as necessary. We assigned automation and manual capabilities as independent variables and set both capabilities to vary between five levels. In addition, we measured the percentage of automation use as the dependent variable. Our main purpose was to investigate whether a mutual dependence exists between calibration and resolution as defined in this study.

1.4 Research Questions and Hypotheses

In this study, we investigated the relationship between calibration and resolution from two different perspectives. First, we investigated what type of resolution would be shown in a situation where users exhibit good calibration by adaptively using automation. Research question one is as follows:

RQ 1: What type of resolution would be shown in a situation where users exhibit good calibration by adaptively using automation?

In this study, using two types of tracking tasks (explained below), we set up a situation where the automation and manual capabilities changed dynamically. Previous studies indicated that when the automation performance is variable, the participants elevated their vigilance against automation performance and facilitated the proper detection of automation errors [8], [10], [11]. In the task situation in this study, we predicted that the participants would elevate their vigilance against variations in the performance of both automation and manual operation and facilitate good calibration. In addition, in such a situation, participants are generally expected to react equally to changes in automation and manual performance when choosing whether to use automation. Therefore, hypothesis I is as follows:

Hypothesis I: Unbiased resolution would appear in a situation where users adaptively use automation.

Second, we divided the participants into two groups, misuse- and disuse-biased groups, and compared their resolution biases. Research question two is as follows:

RQ 2: What type of resolution would be shown by misuse- and disuse-biased users?

In the single-task situation in the current study, we predicted that users who tend to use automation would have more opportunities to recognize changes in automation performance than users who tend to conduct manual operation. Conversely, users who tend to conduct manual operation would have more chance to recognize changes in manual performance than users who tend to use automation. Therefore, hypothesis II is as follows:

Hypothesis II: Misuse-biased users would show more pronounced automation-biased resolution than disuse-biased users would, and disuse-biased users would show more pronounced manual-biased resolution than misuse-biased users would.

2. Experimental Task

We invented two tracking tasks (line and road tasks) as experimental tasks in order to measure calibration and resolution defined in this study (Fig. 5). In the line task, participants track a line that scrolls downward past a circular vehicle. When the vehicle veers off the line, the performance score is reduced according to the operational error. In the road task, participants track a road that scrolls downward past a dot representing a vehicle. When the vehicle hits the edge of the road, the performance score is reduced according to the operational error. The participants were allowed to switch to either the auto mode (operation performed entirely by the system) or the manual mode (operation performed by participants using left and right arrow keys) by pressing a selector on the keyboard.

The circular and dot vehicles in the two tasks are 24 pixels across in diameter. The window scrolling speed is 24 pixels per second in both tasks. In the line task, the line is 5 pixels wide. In the road task, the edge of the road is placed

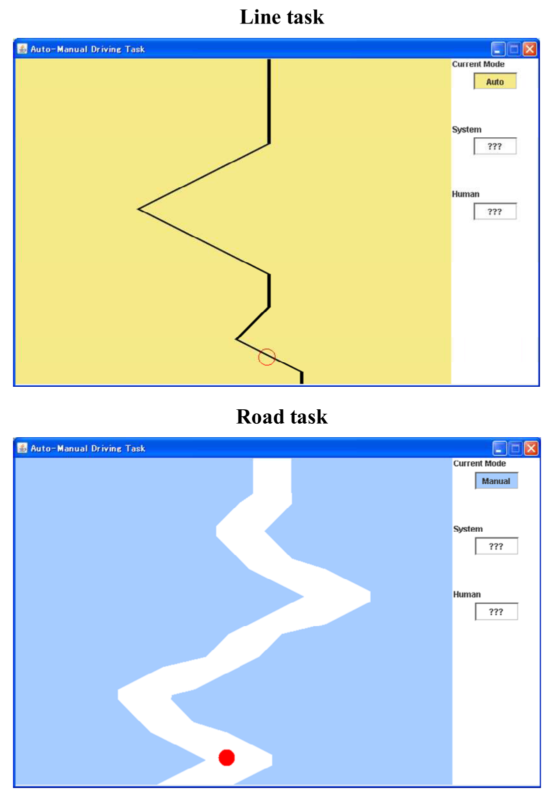

Fig. 5 Line and road tasks.

24 pixels right and left from the invisible center line. Therefore, in the both tasks, the operational errors are designed to occur when the vehicles veer off from the line in the line task or the center line in the road task. When operational errors occur, the task window would be surrounded by a flashing red square frame as error feedback.

In these tasks, we varied the automation capability (Ca) and manual capability (Cm) between five levels: 30, 40, 50, 60, and 70. In this program, system operation in the auto mode and manual operation in the manual mode are normally reflected in the vehicle movements with a sampling rate of 50 Hz. The value of Ca or Cm indicates the percentage how much system or manual operation is reflected in the actual vehicle movements (Fig. 6). For example, when Ca is 30, system operation, which is always perfect, is reflected in only 30% of the vehicle's movements. The other 70% of the system operation is accepted in the experimental task system as no operational command. Therefore, the vehicle does not appropriately track the line or the road in auto mode. In this case, the participants could consider the capability of the auto tracking system to be low. In the same way, when Cm is 30, the manual operation is reflected in only 30% of the vehicle's movements. The other 70% of the manual operation is accepted in the experimental task system as no operational command. Therefore, even if the participant's manual operation is perfect, the vehicle does not appropriately track the line or the road, and the task performance would become low in manual mode. In this case, the participants would experience low vehicle operability.

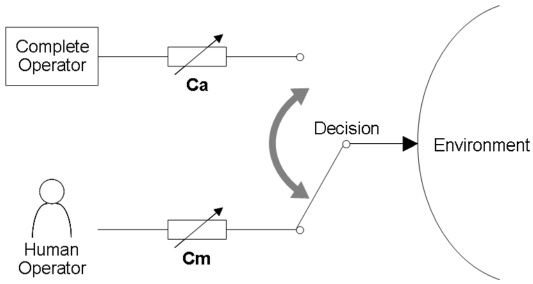

Fig. 6 Experimental manipulation of Ca and Cm in two tracking tasks. System operation in auto mode is always perfect. However, the operation is reflected in only 30% to 70% of the vehicle's movements. In the same way, human operation in the manual mode is reflected in only 30% to 70% of the vehicle's movements. The participants had to switch modes, monitor the vehicle movements in the auto and manual modes, and decide to select the mode that performs better.

Therefore, as Ca or Cm increases, the vehicle becomes more controllable. In contrast, as Ca or Cm decreases, the vehicle becomes less controllable. The participants had to monitor the vehicle movements in the auto and manual modes, and choose the mode that performs better.

The line and road tasks differ in the difficulty of comparing the performance in the auto and manual modes. In the line task, when operational errors occur, the vehicle keeps moving away from the line. Therefore, both the automation and manual performance are clearly visible in terms of the distance from the vehicle to the line. Conversely, in the road task, even when operational errors occur, the dot does not go over the edge of the road but keeps tracking the road in contact with the inner edge. Therefore, it is more difficult to compare the automation and manual performance in this task because differences between the automation and manual performance are not clearly visible. In addition to the research questions, we also investigated the difference in calibration and resolution in the two different task situations.

3. Experiment 1

In the line and road tasks, Ca and Cm were each varied between five levels (30, 40, 50, 60, and 70). As stated earlier, we assumed that when Ca and Cm are equal, the automation and manual operation performance are identical. However, this assumption is not guaranteed in the actual experiment because human control errors are likely to occur and degrade manual performance. Therefore, in Experiment 1, we first measured the auto and manual performances at each value of Ca and Cm. Second, we calculated the formula describing the relationship between Ca and Cm on the basis of the measured auto and manual performances. Finally, the calculated formula in Experiment 1 was used for the analysis in the main experiment, Experiment 2.

3.1 Purpose

The purpose of Experiment 1 was to calculate the formula relating Ca to Cm when the auto and manual performance were identical in each task.

3.2 Method

3.2.1 Participants

One hundred thirty-two university students participated in Experiment 1. Sixty-five of them performed the line task, and the other sixty-seven performed the road task.

3.2.2 Procedure

To measure the manual performance, the participants conducted either the line or road task using only the manual mode. Participants performed a total of 20 trials of each task, consisting of four trials at each Cm value (30, 40, 50, 60, 70). The order of the Cm values was randomized during the task. Each trial lasted for 40 seconds. When one trial ended and the next began, the display showed "Capabilities change" in the center of the screen. At the same time, the number of completed trials among the 20 trials was shown. Before conducting each task, the participants performed one training trial for 40 seconds as practice for manual operation. In the training trial, Cm was set to 100. Throughout the experiment, the Cm values were not displayed on the screen. The participants were required to achieve as high a score as possible. At the end of the task, the task performance score was displayed on the screen.

The auto performance was measured through computer simulations; that is, we measured the auto performance using only the auto mode for each task. For each task, we conducted as many trials at each value of Ca (30, 40, 50, 60, and 70) as there were participants (line: 65 trials each, road: 67 trials each), which can be considered sufficient to verify the auto performance at each value of Ca.

3.3 Results

First, we calculated the manual performance of each participant in each trial. In the line task, we did this by dividing the time the vehicle was on the line by the total time of one trial, 40 seconds. In the road task, we did this by dividing the time the vehicle tracked the road without crashing into the edge of the road by the total time of one trial, 40 seconds. Next we calculated the average manual performance at each value of Cm from 260 data sets (65 participants × four trials) for the line task and 268 data sets (67 participants × four trials) for the road task.

In the same manner, we calculated the average auto performance at each value of Ca from 65 data sets for the line task and 67 data sets for the road task. Finally, we used a linear approximation of the average performance of auto

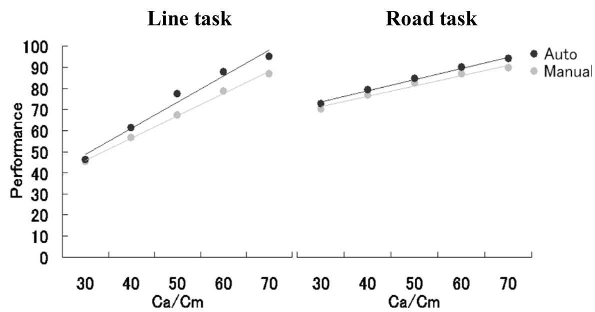

Fig. 7 Average performance of auto and manual modes at each Ca and Cm value, and linear approximation. The x-axis represents Ca or Cm and the y-axis represents the task performance.

and manual mode. Figure 7 shows the average performance of the auto and manual modes at each value of Ca and Cm, and the result of the linear approximation. The auto mode showed slightly higher performance than the manual mode in both tasks.

The calculated approximation formulae for auto and manual performance in each task are as follows:

Line task

Auto mode

$$\text{Performance} = 1.235Ca + 11.766 \tag{1}$$

Manual mode

$$\text{Performance} = 1.049Cm + 14.521 \tag{2}$$

Road task

Auto mode

$$\text{Performance} = 0.530Ca + 57.315 \tag{3}$$

Manual mode

$$\text{Performance} = 0.488Cm + 56.389 \tag{4}$$

Finally, for each task, we calculated the formula describing the relationship between Ca and Cm from the approximation formulae when the auto and manual performance are identical.

Line task

$$Cm = 1.177Ca - 2.626 \tag{5}$$

Road task

$$Cm = 1.086Ca + 1.897 \tag{6}$$

3.4 Discussion

If the participants performed as well as the automation system did, the performance scores of the auto and manual modes at each value of Ca and Cm would be equal in both the line and road tasks. However, the auto mode showed higher performance than the manual mode did in both tasks. Human factors such as control errors and fatigue may cause this disadvantage in the manual operation.

The performance scores were lower in the line task than in the road task, especially at the lower values of Ca and Cm. When operational errors occur, the circular vehicle keeps moving away from the line in the line task; but the dot vehicle stays within the road in the road task. Therefore, it is more difficult to recover from the operational errors in the line task than in the road task.

In Experiment 2, we investigated the research questions, using the relationship formula for Ca and Cm obtained in Experiment 1.

4. Experiment 2

4.1 Purpose

The purpose of Experiment 2 was to investigate RQ 1 and 2.

4.2 Method

4.2.1 Participants

Twenty-seven university students participated in Experiment 2. They performed both the line and the road tasks, and the order of the tasks was counterbalanced among the participants. Four participants were excluded from analysis because of machine trouble.

4.2.2 Procedure

The participants performed both the line and road tasks using the auto and manual modes. For each task, we conducted 25 trials. Each trial consisted of one of 25 combinations of 5 (Ca: 30, 40, 50, 60, 70) × 5 (Cm: 30, 40, 50, 60, 70). Throughout each task, participants experienced all the 25 combinations. The order of the Ca and Cm values was randomized during the task. Each trial lasted for 40 seconds. When one trial ended and the next began, the display showed "Capabilities change" in the center of the screen. At the same time, the number of completed trials among the 25 trials was shown. Before conducting each task, the participants performed two training trials for 40 seconds each as practice for switching between the auto and manual modes. In the first training trial, Ca was set to 70 and Cm was set to 30; in the second training trial, Ca was set to 30 and Cm was set to 70. Throughout the experiment, the Ca and Cm values were not displayed on the screen. Therefore, the participants were not informed of the values. The participants were required to achieve as high a score as possible for each task, adaptively selecting either the auto or manual mode. At the end of each task, the task performance score was displayed on the screen.

4.3 Results

First, the individual percentage of auto mode use for each combination of Ca (5 levels) × Cm (5 levels) was recorded for each task. In particular, the percentage of auto mode use in each trial was calculated by dividing the time each participant used auto mode by the total time of one trial, 40

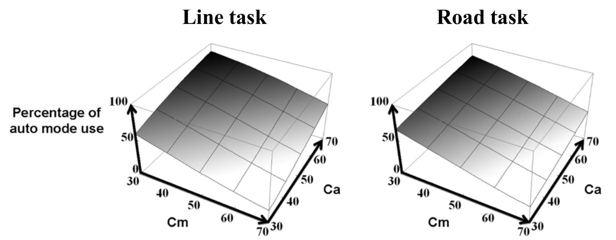

Fig. 8 Predicted curve for each task. The x-axis represents Cm, the y-axis represents Ca, and the z-axis represents the percentage of auto mode use.

seconds. Next, we calculated the average percentage of auto mode use by the 23 participants for each combination of Ca (5 levels) × Cm (5 levels). Finally, we fitted the logistic curve to the average percentages of auto mode use at the 25 data points. The predicted percentages of auto mode use are as follows:

Line task

$$\text{Percentage of auto mode use} = \frac{100}{1 + \exp -(0.505 + 0.042Ca - 0.046Cm)} \tag{7}$$

Road task

$$\text{Percentage of auto mode use} = \frac{100}{1 + \exp -(1.317 + 0.022Ca - 0.044Cm)} \tag{8}$$

Figure 8 shows the predicted curve for each task. We used the Hosmer-Lemeshow test to assess the goodness of fit of the predicted curves to the observed average percentage of auto mode use. We found that the test was significant in neither the line (p = .89) nor the road (p = .97) tasks, indicating that the logistic curves described the data well. In addition, Fig. 9 shows the cross-sectional figures of the predicted curve and the observed average percentage of auto mode use in each task. Figure 9 also shows that only a slight deviation appears between the predicted curve and the observed average percentage of auto mode use.

4.3.1 Evaluation of Calibration and Resolution

To evaluate the calibration for each task, we used the predicted percentage of auto mode use when Ca and Cm are both 50 as a representative point. The graphical position of this point is located at the center of the predicted surface (Fig. 8). The efficiency of auto mode use differs on either side of the central position, as shown in Fig. 2. Therefore, when the auto mode was used appropriately, the predicted percentage of auto mode use would decrease in front of the central position when manual operation is effective. In contrast, the predicted percentage of auto mode use would increase behind the central position when auto mode use is effective. In this case, the predicted percentage of auto mode use at the central position would shift neither upward nor downward, but would appear approximately 50%. Thus,

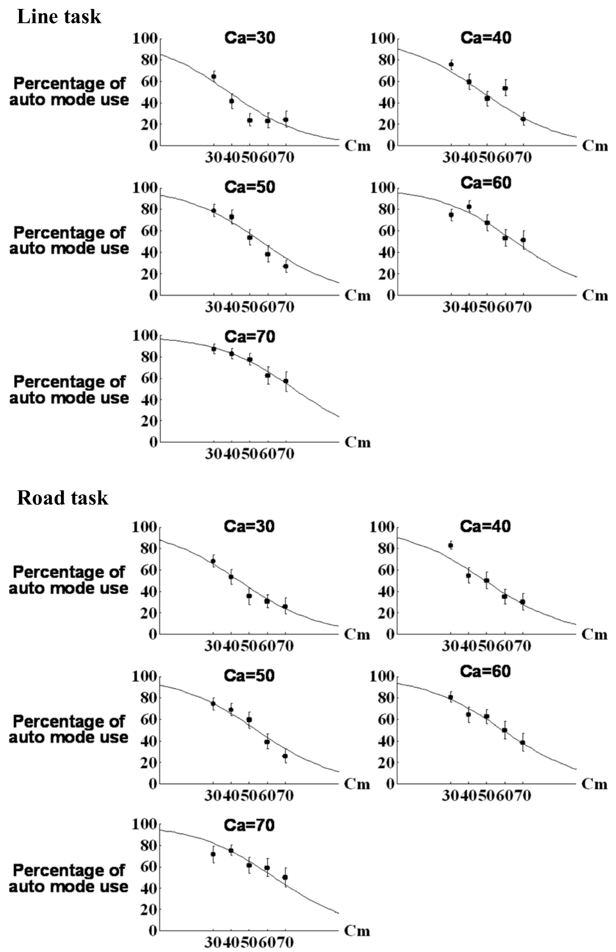

Fig. 9 Cross-sectional figures of predicted curve and average percentages of auto mode use in each task. The curve represents the logistic regression curve for each value of Ca, the dots represent the observed average percentage of auto mode use for each combination of Ca (5 levels) and Cm (5 levels), and the error bars represent the standard errors.

auto mode usage of approximately 50% at this representative point indicates that participants' calibration is neither misuse- nor disuse-biased but represents good calibration. However, automation use of more than 50% at this representative point indicates that participants tend to use the auto mode even when it is not effective, and thus become misuse-biased. Conversely, automation use of less than 50% at this representative point indicates that participants tend to use manual operation even when using the auto mode is effective and thus become disuse-biased.

To evaluate the resolution for each task, we compared the odds ratios of Ca and Cm calculated from the logistic regression formula based on Hosmer and Lemeshow [20]. The odds ratio of Ca represents the increase rate of auto mode use with changes in Ca, and the odds ratio of Cm represents the decrease rate of auto mode use with changes in Cm. To compare these ratios, we calculated their product and evaluated whether the product exceeds one or falls below one. A value of one indicates that the percentage of auto mode use changes equally with changes in Ca and Cm. It shows that

Table 1 Predicted percentage of auto mode use when Ca and Cm are both 50, odds ratios (ORs) of Ca and Cm, and product of odds ratios of Ca and Cm for each task.

Task	Percentage of auto mode use	Percentage* of auto mode use	OR of Ca	OR of Cm	OR of Cm*	Product of ORs of Ca and Cm	Product of ORs of Ca and Cm*
Line	57.245	50.055	1.522	0.629	0.579	0.958	0.882
Road	54.603	47.652	1.248	0.638	0.614	0.797	0.767

participants react equally to changes in Ca and Cm, meaning that their resolution is unbiased. However, a value exceeding one indicates that the percentage of auto mode use changes greatly with changes in Ca than with those in Cm. It shows that participants react more sensitively to changes in Ca, meaning that their resolution tends to be automation-biased. Conversely, a value of less than one shows that participants react more sensitively to changes in Cm, meaning that their resolution is manual-biased.

4.3.2 Investigation of RQ 1 and 2

Table 1 shows the predicted percentage of auto mode use when Ca and Cm are both 50, the odds ratios of Ca and Cm, and the product of the odds ratios of Ca and Cm for each task. The percentage of auto mode use and the odds ratios of Ca and Cm were calculated from the logistic regression formulae (7) and (8). However, these formulae do not reflect the differences between auto and manual performance observed in Experiment 1. Therefore, we corrected them using the relationship formulae obtained in Experiment 1. The correction and calculation methods are described in Appendix A.

It is guaranteed that when Ca and Cm* (the corrected value of Cm) are equal in the corrected logistic regression, the auto and manual modes exhibit identical task performance. In Table 1, the percentage* of auto mode use was calculated by substituting 50 for Ca and Cm* in the corrected logistic regression formula for each task. In addition, the odds ratio of Cm* was also calculated from the corrected logistic regression formula in each task, representing the degree of change in the percentage* of auto mode use with a change in Cm*. In this study, we used the percentage* of auto mode use to evaluate the calibration and the product of the odds ratios of Ca and Cm* to evaluate the resolution. The corrected logistic regression formulae are described in Appendix B.

First, we investigated RQ 1. For the calibration, the percentage* of auto mode use settled approximately 50% in each task. As we predicted, the result showed that the participants' tendency to use the auto mode was neither misuse- nor disuse-biased but exhibited good calibration. Conversely, for the resolution, the product of the odds ratios of Ca and Cm* fell below one for each task. This indicates that the participants showed manual-biased resolution; that is, they reacted more sensitively to changes in Cm* than to changes in Ca. The participants reacted 1.133 (= 1/0.882) and 1.303 (= 1/0.767) times more sensitively to changes in Cm* than to changes in Ca in the line and road tasks, respectively. Therefore, hypothesis I was rejected. In addition, the participants showed more pronounced manual-biased resolution in the line task than in the road task.

Second, we investigated RQ 2. We divided the participants into misuse- and disuse-biased groups according to the median average of the percentage of auto mode use in each task to compare the resolution biases in both groups. In each task, we excluded the participant whose percentage of auto mode use was the median average from analysis. Therefore, each group consisted of eleven participants in each task. For each group and each task, the average percentage of auto mode use was recorded for each combination of Ca (5 levels) × Cm (5 levels). Next, we fitted the logistic curve to the 25 data points. The predicted percentages of auto mode use are as follows:

Line task

Misuse-biased group

Percentage of auto mode use

$$= \frac{100}{1 + \exp -(0.249 + 0.047Ca - 0.034Cm)} \tag{9}$$

Disuse-biased group

Percentage of auto mode use

$$= \frac{100}{1 + \exp -(0.522 + 0.045Ca - 0.062Cm)} \tag{10}$$

Road task

Misuse-biased group

Percentage of auto mode use

$$= \frac{100}{1 + \exp -(1.696 + 0.031Ca - 0.047Cm)} \tag{11}$$

Disuse-biased group

Percentage of auto mode use

$$= \frac{100}{1 + \exp -(1.080 + 0.017Ca - 0.048Cm)} \tag{12}$$

The Hosmer-Lemeshow test was not significant in either the misuse- (p = .94) or the disuse- (p = .78) biased groups in the line task, or in the misuse- (p = .95) or disuse- (p = .93) biased groups in the road task, indicating that the logistic curves described the data well.

Table 2 shows the predicted percentage of auto mode use when Ca and Cm are both 50, the odds ratios of Ca and Cm, and the product of the odds ratios of Ca and Cm in the misuse- and disuse-biased groups. The percentage of auto mode use and the odds ratios of Ca and Cm were calculated

Table 2 Predicted percentage of auto mode use when Ca and Cm are both 50, odds ratios (ORs) of Ca and Cm, and product of odds ratios of Ca and Cm in misuse- and disuse-biased groups.

Task	Group	Percentage of auto mode use	Percentage* of auto mode use	OR of Ca	OR of Cm	OR of Cm*	Product of ORs of Ca and Cm	Product of ORs of Ca and Cm*
Line	Misuse	71.054	66.482	1.600	0.711	0.669	1.138	1.071
	Disuse	42.192	33.070	1.578	0.535	0.479	0.845	0.757
Road	Misuse	71.490	65.156	1.372	0.623	0.598	0.856	0.821
	Disuse	38.319	31.499	1.189	0.616	0.590	0.732	0.702

from the logistic regression formulae (9)–(12). However, the percentage* of auto mode use and the odds ratio of Cm* were calculated from the corrected logistic regression formulae in Appendix B.

For the resolution, we compared the products of the odds ratios of Ca and Cm* in the misuse- and disuse-biased groups for each task. For both tasks, the products of the odds ratios of Ca and Cm* were smaller in the disuse-biased group than in the misuse-biased group. That is, the disuse-biased users showed more pronounced manual-biased resolution than the misuse-biased users did. This result supports hypothesis II. In addition, both misuse- and disuse-biased participants showed more pronounced manual-biased resolution in the line task than in the road task.

4.4 Discussion

First, for RQ 2, as stated in hypothesis II, the disuse-biased group showed more pronounced manual-biased resolution than the misuse-biased group did. Therefore, we confirmed that a mutual dependence exists between calibration and resolution. Second, for RQ 1, as we predicted on the basis of previous studies [8], [10], [11], in a task situation where both automation and manual performance are variable, the participants generally tended to exhibit good calibration by using automation adaptively. However, contrary to hypothesis I, although they used automation adaptively, they showed not unbiased resolution but manual-biased resolution.

5. General Discussion

5.1 Manual-Biased Resolution

For RQ 1, we investigated what type of resolution would generally appear when users exhibit good calibration by adaptive use of automation. We found that, contrary to hypothesis I, participants in such a situation showed not unbiased resolution but manual-biased resolution.

First, to examine the calibration, we set up a situation in this study where the automation and manual capabilities are both variable. Participants elevated their vigilance against changes in both automation and manual performance, as in previous studies [8]–[11]. Consequently, they could adaptively select whether to use automation or perform manual operation without falling into automation misuse or disuse.

Second, when we examined the resolution, participants chose whether to use automation or perform manual operation while reacting more sensitively to changes in manual performance than those in automation performance. This result is explained in terms of human cognitive capacities and situation awareness. Human cognitive capacities are limited. In addition, previous studies of situation awareness showed that people have superior awareness during active human monitoring (monitoring situations while manually conducting a task) than during passive automation monitoring (monitoring situations while observing an automated operation) [21]–[23]. In this study, participants were required to evaluate both automation and manual performance to maximize their task performance. However, they might not be able to consider both changes simultaneously because of the limitations on human cognitive capacities; that is, it was not possible for participants to memorize the performance in one mode and compare the memorized performance to the actual performance in the other mode. Consequently, they tended to select automation use on the basis of the performance of only one mode. Under this constraint, we assume that our participants adopted manual-biased resolution because active human monitoring provides better situation awareness than passive automation monitoring, and they chose to use automation adaptively.

5.2 Relationship between Calibration and Resolution

For RQ 2, we investigated what type of resolution would be shown by misuse- and disuse-biased users. Theoretically, calibration and resolution are considered to be independent human behaviors. However, the results of our experiments confirmed that disuse-biased users showed more pronounced manual-biased resolution than misuse-biased users did. We suggest two possible explanations for this behavioral tendency.

One is that trust might act as a link between calibration and resolution. Previous studies showed that a relationship exists between human use of automation and human trust in automation [1]–[4], [6]. Users who trust automation tend to overuse it, whereas those who distrust automation tend to underuse it. In a single-task situation, we assume that disuse-biased participants distrusted automation and tended to conduct manual operation, and thus allocated more attention to manual operation than the misuse-biased participants. Therefore, disuse-biased participants might be able to recognize changes in the manual performance and show more pronounced manual-biased resolution than the misuse-biased participants.

Another possibility is individual differences in the ability to recognize changes in the automation performance. Madhavan and Phillips [18] showed that there are individual differences in the ability to change the decision to use automation with changes in automation performance. We assume that the disuse-biased participants might have lower levels of this ability than the misuse-biased participants. As a result, the disuse-biased participants might react more sensitively to changes in the manual performance than to changes in the automation performance when they decide whether to use automation.

5.3 Influence of Task

We used two types of tracking tasks in this study. It was more difficult to evaluate the automation and manual performance in the road task than in the line task. Our experimental results showed that participants exhibited more pronounced manual-biased resolution in the road task than in the line task. It is possible that in the road task, participants might evaluate the manual performance based on the controllability of the dot to compensate for the difficulty in visually evaluating the performance.

Metcalfe and Greene [24] experimentally investigated the nature of human judgment of agency (JOA), i.e., self-judgment of control of their own actions, using a computer. In their study, the participants had to control the cursor on the computer screen with a mouse and catch a target object falling from the top of the screen. The controllability of the cursor was manipulated at various levels, and participants could not control the cursor as they thought they could. The results of their experiments showed that participants could appropriately evaluate how much their manipulation affected the actual movement of the cursor; that is, they appropriately judged the degree of controllability of the cursor in their JOA. In our experiment, it was more difficult to evaluate both the automation and manual performance in the road task than in the line task. In such a situation, participants might be able to evaluate the controllability of the dot vehicle and compensate for the difficulty in performance evaluation. Therefore, they showed more pronounced manual-biased resolution in the road task than in the line task.

6. Conclusion

In this study, we investigated human use of automation based on the automation and manual performance using the concepts of calibration and resolution. This investigation was possible because we used an innovative performance-based analysis. Our experiments showed that a mutual relationship exists between calibration and resolution. Disuse-biased users showed more pronounced manual-biased resolution than misuse-biased users did. There are two possible explanations for this phenomenon: one is that trust might act as a link between calibration and resolution, and the other is that individual differences in the ability to recognize changes in the automation performance are related to the calibration bias.

Furthermore, in a situation where good calibration is exhibited, when users determine whether to use automation, they do not react to changes in the automation and manual performance evenly; rather, they react more sensitively to changes in manual performance. We assume that this tendency in human behavior arises from the limitations of human cognitive capacities and the superiority of active human monitoring to passive monitoring.

In this study, we focused on human reactions to automation in a changing environment; therefore, we did not ask participants to rate their trust during the task to retain task continuity. In future research, the relationship among automation capability, subjective trust in automation, and automation usage needs to be investigated. In addition, the relationship among manual capability, subjective self-confidence in manual operation, and automation usage also needs to be investigated.

References

[1] J.D. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," Ergonomics, vol.35, no.10, pp.1243–1270, Oct. 1992.
[2] J.D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," Int. J. Human-Computer Studies, vol.40, no.1, pp.153–184, Jan. 1994.
[3] S. Lewandowsky, M. Mundy, and G.P.A. Tan, "The dynamics of trust: Comparing humans to automation," J. Experimental Psychology: Applied, vol.6, no.2, pp.104–123, June 2000.
[4] N. Moray, T. Inagaki, and M. Itoh, "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks," J. Experimental Psychology: Applied, vol.6, no.1, pp.44–58, March 2000.
[5] B.M. Muir, "Trust in automation: Part 1. Theoretical issues in the study of trust and human intervention in automated systems," Ergonomics, vol.37, no.11, pp.1905–1922, Nov. 1994.
[6] B.M. Muir and N. Moray, "Trust in automation. part II. experimental studies of trust and human intervention in a process control simulation," Ergonomics, vol.39, no.3, pp.429–460, March 1996.
[7] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," Human Factors, vol.39, no.2, pp.230–253, June 1997.
[8] R. Parasuraman, R. Molloy, and I.L. Singh, "Performance consequences of automation-induced 'complacency'," The Int. J. Aviation Psychology, vol.3, no.1, pp.1–23, Jan. 1993.
[9] R. Molloy and R. Parasuraman, "Monitoring an automated system for a single failure vigilance and task complexity effects," Human Factors, vol.38, no.2, pp.311–322, June 1996.
[10] I.L. Singh, R. Molloy, and R. Parasuraman, "Individual differences in monitoring failures of automation," J. General Psychology, vol.120, no.3, pp.357–373, July 1993.
[11] I.L. Singh, R. Molloy, and R. Parasuraman, "Automation-induced monitoring inefficiency: Role of display location," Int. J. Human-Computer Studies, vol.46, no.1, pp.17–30, Jan. 1997.
[12] R. Parasuraman and D.H. Manzey, "Complacency and bias in human use of automation: An attentional integration," Human Factors, vol.52, no.3, pp.381–410, June 2010.
[13] H.P. Beck, M.T. Dzindolet, and L.G. Pierce, "Automation usage de-

cisions: Controlling intent and appraisal errors in a target detection task," Human Factors, vol.49, no.3, pp.429–437, June 2007.
[14] H.P. Beck, J.B. McKinney, M.T. Dzindolet, and L.G. Pierce, "Effects of human-machine competition on intent errors in a target detection task," Human Factors, vol.51, no.4, pp.477–486, Aug. 2009.
[15] M.T. Dzindolet, L.G. Pierce, H.P. Beck, and L.A. Dawe, "The perceived utility of human and automated aids in a visual detection task," Human Factors, vol.44, no.1, pp.79–94, Jan. 2002.
[16] M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce, and H.P. Beck, "The role of trust in automation reliance," Int. J. Human-Computer Studies, vol.58, no.6, pp.697–718, June 2003.
[17] J.D. Lee and K.A. See, "Trust in automation: Designing for appropriate reliance," Human Factors, vol.46, no.1, pp.50–80, Jan. 2004.
[18] P. Madhavan and R.R. Phillips, "Effects of computer self-efficacy and system reliability on user interaction with decision support systems," Computers in Human Behavior, vol.26, no.2, pp.199–204, March 2010.
[19] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," Int. J. Human-Computer Studies, vol.58, no.6, pp.719–735, June 2003.
[20] D.W. Hosmer and S. Lemeshow, Applied logistic regression, Wiley & Sons, New York, 1989.
[21] M.R. Endsley and E.O. Kiris, "The out-of-the-loop performance problem and level of control in automation," Human Factors, vol.37, no.1, pp.381–394, June 1995.
[22] U. Metzger and R. Parasuraman, "The role of the air traffic controller in future air traffic management: An empirical study of active control versus passive monitoring," Human Factors, vol.43, no.4, pp.519–528, Dec. 2001.
[23] R. Parasuraman, M. Mouloua, and R. Molloy, "Effects of adaptive task allocation on monitoring of automated systems," Human Factors, vol.38, no.4, pp.665–679, Dec. 1996.
[24] J. Metcalfe and M.J. Greene, "Metacognition of agency," J. Experimental Psychology: General, vol.136, no.2, pp.184–199, May 2007.

Appendix A

Correction method of logistic regression formula

1) Calculate the linear approximation formula of each auto and manual performance as in Experiment 1.
2) Calculate the relation formula of the auto and manual performances from the linear approximation formulae as in Experiment 1.

$$Cm = \mathrm{x}Ca + \mathrm{y} \tag{A$\cdot$1}$$

3) Calculate the logistic regression formula as in Experiment 2.

$$\text{Percentage of auto mode use} = \frac{100}{1 + \exp -(\beta_0 + \beta_1 Ca + \beta_2 Cm)} \tag{A$\cdot$2}$$

4) Transform the relation formula of the auto and manual performances to the corrected formula. In particular, Ca in the relation formula obtained in Experiment 1 is replaced with Cm*.

$$Cm = \mathrm{x}Cm* + \mathrm{y} \tag{A$\cdot$3}$$

5) Assign the corrected formula to the logistic regression formula. By this correction, when the values of Ca and Cm* are equal, the auto performance with Ca and the manual performance with x*Cm**+y are identical.

$$\text{Percentage* of auto mode use} = \frac{100}{1 + \exp -(\beta_0 + \beta_1 Ca + \beta_2(\mathrm{x}Cm* + \mathrm{y}))} \tag{A$\cdot$4}$$

Calculation method of odds ratio

Odds ratios of Ca and Cm are calculated from the logistic regression formula:

$$\text{Odds ratio of } Ca = \exp(10\beta_1) \tag{A$\cdot$5}$$

$$\text{Odds ratio of } Cm = \exp(10\beta_2) \tag{A$\cdot$6}$$

Odds ratios of Cm* are calculated from the corrected logistic regression formula:

$$\text{Odds ratio of } Cm* = \exp(10\beta_2 \mathrm{x}) \tag{A$\cdot$7}$$

Appendix B

Corrected logistic regression formulae in Experiment 2

Line task

Percentage* of auto mode use

$$= \frac{100}{1+\exp-(0.505+0.042Ca-0.046(1.177Cm*-2.626))} \tag{A$\cdot$8}$$

Road task

Percentage* of auto mode use

$$= \frac{100}{1 + \exp -(1.317 + 0.022Ca - 0.044(1.086Cm* + 1.897))} \tag{A$\cdot$9}$$

Line task

Misuse-biased group

Percentage* of auto mode use

$$= \frac{100}{1 + \exp -(0.249 + 0.047Ca - 0.034(1.177Cm* - 2.626))} \tag{A$\cdot$10}$$

Disuse-biased group

Percentage* of auto mode use

$$= \frac{100}{1 + \exp -(0.522 + 0.045Ca - 0.062(1.177Cm* - 2.626))} \tag{A$\cdot$11}$$

Road task

Misuse-biased group

Percentage* of auto mode use

$$= \frac{100}{1 + \exp -(1.696 + 0.031Ca - 0.047(1.086Cm* + 1.897))} \tag{A$\cdot$12}$$

Disuse-biased group

Percentage* of auto mode use

$$= \frac{100}{1 + \exp -(1.080 + 0.017Ca - 0.048(1.086Cm * +1.897))} \tag{A·13}$$

Akihiro Maehigashi received the M.A. degree from Nagoya University in 2010. He became a doctor course student in 2010 and a Researcher in 2013 in the Graduate School of Information Science at Nagoya University.

Kazuhisa Miwa received the B.S. degrees in Applied Physics from Nagoya University in 1984. He received the Ph.D. degree in Information Science from Nagoya University in 1989. He stayed in CMU as a visiting scholar from 1991 to 1992. He is currently a professor of Nagoya University. His research interests include cognitive science and artificial intelligence.

Hitoshi Terai received the Ph.D. degree from Nagoya University in 2006. He is currently a Designated Associate Professor in the Graduate School of Information Science at Nagoya University: Core Research for Evolutionary Science and Technology, Japan Science and Technology Agency.

Kazuaki Kojima received the Ph.D. degree from Nagoya University in 2007. From 2007 to 2009, he was a Research Associate, and from 2009 to 2013, he was an Assistant Professor in Faculty of Human Sciences, Waseda University. He is currently an Assistant Professor in Learning Technology Laboratory, Teikyo University.

Junya Morita received the Ph.D. degree from Nagoya University in 2006. He is currently an Assistant Professor in the School of Knowledge Science at Japan Advanced Institute of Science and Technology.