Effects of Saliency of an Agent's Input Information on Estimation of Mental States toward the Agent

Yuki Ninomiya
Institute of Innovation for Future Society
Nagoya University
Aichi, Japan
0000-0002-6032-8003

Asaya Shimojo
KONICA MINOLTA,Inc.
Tokyo, Japan
email:asaya.shimojo
@konicaminolta.com

Shota Matsubayashi
Institute of Innovation for Future Society
Nagoya University
Aichi, Japan
email:matsubayashi.shota.v0
@f.mail.nagoya-u.ac.jp

Hitoshi Terai
Faculty of Humanity-Oriented Science and Engineering
KINDAI University
Fukuoka, Japan
email:teraihitoshi@gmail.com

Kazuhisa Miwa
Graduate School of Informatics
Nagoya University
Aichi, Japan
email:miwa@is.nagoya-u.ac.jp

***Abstract*—Humans predict the behaviors of an autonomous agent by estimating its mental state via anthropomorphization of the agent. This paper examines the effect of the saliency of input information used by an agent on user estimation of the agent's mental state. The results demonstrate that observers can correctly estimate the mental states of agents whose input information has both high and low saliency. However, we found that observers face difficulties when asked to report their estimations verbally. This suggests that a discrepancy exists between the estimation of the agent's mental state and the user's verbal reporting.**

Keywords; agent, goal inference, theory of mind.

I. Introduction

In recent years, a variety of different autonomous agents have been developed and used in practical applications. In such situations, users are required to predict and understand the behavior of autonomous agents. Here, humans may attempt to identify the cause of behavior by anthropomorphizing the object and estimating its mental state by, for example, wondering what the vacuum cleaner is having trouble with [1]. If there is something observable, e.g., an obstacle, the cause can be identified easily. However, if the cause is difficult to recognize to just at first glance, e.g., a slippery floor, it is impossible to estimate what the agent is struggling with. In this paper, we examine the estimation of an agent's mental state to predict and understand an agent's behaviors.

Projecting a mental state onto an agent is useful relative to predicting and understanding the agent's behavior [1]. Reference [1] explained that knowledge about the general human being serves as an easily accessible base for estimating the mental states and characteristics of an unknown agent. In other words, humans predict the behavior of an unknown non-human agent by projecting common human mental states, e.g., beliefs and desires, onto the non-human agent.

Many studies have demonstrated the effectiveness of estimating the mental states of robots and machines. For example, it has been shown that placing robotic eyes on automated vehicles facilitates communication between automated vehicles and pedestrians [2].

However, inaccurately estimating the mental state of such agents can lead to serious accidents [2] [3]. Thus, to allow humans to identify an agent's mental state, it is necessary to clarify how humans correctly estimate agent's mental states.

To correctly estimate the mental state of a target, e.g., its goals or intentions, it is necessary to accurately recognize two types of information, environment as a situational constraint and behavior [4] [5]. In terms of estimating the mental state of an autonomous agent, the environment corresponds to the input information used by the agent, and behavior corresponds to the agent's output. In other words, it is important to correctly recognize what the agent is using as input information and accurately estimate the mental state from the agent's output. For example, consider a situation where an agent moves away from an enemy, as illustrated in Figure 1. Here, the behavior corresponds to agent's movement to the lower left, and the environment corresponds to the presence of the enemy to the agent's upper right. In such a case, the observer estimates that the agent's purpose for moving to the lower left is to escape the enemy.

In this study, we focused on saliency as a factor affecting the accuracy of estimating the mental state of agents. Saliency is a property of how attention-grabbing a stimulus is compared to its surroundings, and salient stimuli or events cause bias of attention [6]. Humans do not perceive all given information equally but distinguish between figure and ground and recognize figure information preferentially [7]. Visual perception studies have demonstrated that more salient features are more likely to be recognized as figure [8]. In addition, high saliency is sometimes taken as an indicator of the ease of recognition as a figure. Although the concepts of figure and ground are related to perception, these concepts are also discussed in

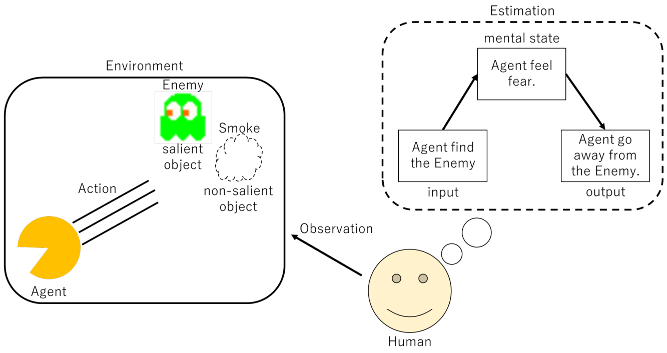

Figure 1. Estimating an agent's mental state.

higher cognitive domains, e.g., reasoning and problem-solving [9].

Thus, the saliency of information may contribute to the ease of recognizing that information when humans estimate the mental state of an agent. In other words, if the saliency of the input information is low, correct estimation of the mental state is likely to be disturbed. Consider the example shown in Figure 1. Here, the agent may not be escaping from the enemy, which is highly salient for the observer, the agent may actually want to escape from the smoke, which is considerably less salient for the observer. If the observer is unaware of the less salient information and only the smoke is present in the environment, it will fail to estimate the agent's mental state or predict its behavior accurately.

This paper examines the effect of the saliency of an agent's input information on the accuracy of the estimation of its mental state. If the saliency of the agent's input information is high, the observer will be more likely to pay attention to that information; thus, the agent's mental state will be estimated correctly. In contrast, if the saliency of the agent's input information is low, observers will be less likely to pay attention to that information; thus, the agent's mental state will be estimated incorrectly. In addition, we examine a case in which the agent uses both high- and low-saliency information as inputs concurrently. In this case, the presence of the high-saliency information can cause the low-saliency information to be neglected by focusing of the high-saliency information. To support this, a previous study found that directing attention to salient information inhibits problem solving that can be achieved by directing attention to less salient information [10]. Thus, we consider the following Research Question (RQ).

RQ1: Can participants (i.e., observers) correctly estimate an agent's mental state even when the agent utilizes less salient information?

We also investigate whether observers can verbally report the information on which they focus to estimate the agent's mental state. Previous research has shown that verbalizing thoughts promotes further focus on information that is easy to pay attention to, and as a result, other information is more likely to be ignored [10] [11]. This finding suggests that verbal reporting may lead to a focus on highly salient information and to ignoring less salient information. In other words, even if the observers can estimate mental states by focusing on the correct information, it may be difficult for them to report less salient information. Thus, we also consider RQ2.

RQ2: Is there a discrepancy between the results of the estimation of mental state and verbal reports?

II. Method

In this section, we describe the experiments to validate the two RQs.

A. Participants

The participants were 108 Japanese university students (N_{female} = 19, N_{male} = 91, M_{age} = 20.23, SD_{age} = 0.88).

B. Procedure

The experiment was conducted using an online environment. The experiment program was created using jspsych [12]. The experimental task comprised three phases, i.e., an observation phase, an estimation phase, and a verbal reporting phase.

In the observation phase, the participants observed a video of an agent moving through a maze (Figure 2(a)). While moving through the maze, the agent changed its speed and color in four steps according to the surrounding environment (Figure2(c)). Here, the number of enemies (zero, one, two, three or more) and the number of escape routes (four or more, three, two, or one) were used as the input information to determine the output (i.e., the speed and color). The color is as shown in Figure2(c), and the speed was quickened in four steps depending on the degree of fear. Enemies are placed as objects in the maze and exist as figures, while the routes require attention to the background information, i.e., the ground information. Thus, information regarding the number of enemies is considered to have higher saliency than the number of escape routes.

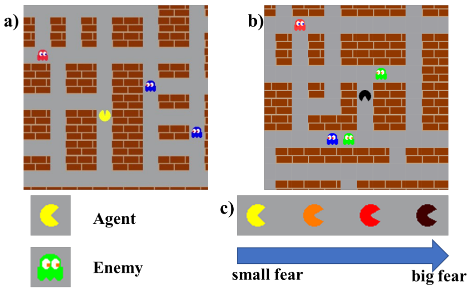

Figure 2. (a) Example screenshot of the observation phase and (b) an evaluation image in the estimation phase. (c) The output (i.e., the color) changes in response to the surrounding environment.

The agents and enemies were created based on PAC-MAN (BANDAI NAMCO Entertainment Inc., 1980), which is a common Japanese video game. The similarity of motion and morphology to humans is a factor that facilitates anthropomorphizing agents [1]. PAC-MAN has a mouth, an organ morphologically similar to that of humans, that opens and closes in a motion that is similar to that of humans; thus, PAC-MAN is easily anthropomorphizable. As a result, this character is a natural target for the estimation of mental states.

Prior to performing the observation phase, the participants were told that the agent was functionally capable of feeling fear and changing its behavior in response to the environment. The participants were then told, “After the observation, you will be asked to answer in which situations the agent felt fear.” The video observed by the participants showed the agent moving through a maze with an enemy from the start to the goal. To experience the full combination of the number of enemies, this set of videos presented 16 patterns, i.e., the number of enemies (four patterns) and the number of escape routes (four patterns).

In the subsequent estimation phase, a screenshot of the experimental video (Figure 2(b)) was displayed to the participants. Then, the participants answered the following question using a seek bar to rate their level of fear on a scale from 0–100: “How scared the agent feels in this situation?”. For each of the 16 patterns, the evaluation stimuli consisted of the number of enemies (four) times the number of escape routes (four) presented four times, i.e., 64 patterns in total.

After the estimation phase, we examined whether the participants could verbally report their estimation of the mental state. Here, the participants responded to which situation the agent felt fear by completing the following if-then sentence: “If ____, the agent would feel fear.” Here, the participants were able to describe as many rules as they could think of.

C. Experimental design

To examine the effect of the saliency of the input information on the estimation of the mental state, we prepared an enemy condition and a route condition. In the enemy condition, the agent changed its output (i.e., its speed and color) using the number of enemies as input, which is highly salient information. For the route condition, the agent changed its output by taking the number of escape routes, which is less salient information, as input. We also set the enemy-path condition, where the agent uses both the low- and high-saliency information as inputs. This condition was designed so that the same weight was used between the number of enemies and the number of paths when determining the output. As a control condition, we added a condition that returns a random output for the input information. We predicted that if the participants could estimate the agent’s mental state correctly, then their mental state estimation was more likely to rely on the information that was used as input the agent in the other conditions than in the control condition.

III. Results

To examine RQ1, we analyzed the participants’ responses in the estimation phase to determine whether they estimated the agent’s mental state by focusing on the correct information. First, the following multiple regression equation (1) was calculated for each participant using the environmental information (number of enemies and routes) in the evaluation images as explanatory variables and the participant’s fear rating as the explained variable. Here, the partial regression coefficients for the number of enemies and escape routes were estimated using the maximum likelihood estimation method.

$$\begin{aligned} Fear =& \beta_{enemy} * number_of_enemies + \\ & \beta_{escaperoute} * number_of_routes + e \end{aligned} \tag{1}$$

We used β as a measure of how much attention the participants paid to each type of information when estimating the agent’s mental state. To examine whether the participants were estimating the correct mental state under each condition, we compared the β value for each condition to that of the control condition. This analysis demonstrated that for all conditions, the β values for the only input information used by the agents were higher than those in the control condition (Figure 3). Note that no significant differences were found for input information not used by the agent. Specifically, there was an interaction in the enemy and route conditions(enemy: $F(1, 53) = 68.11, p < .001$, route: $F(1, 53) = 8.92, p < .01$), but not in the enemy route condition($F(1, 52) = 1.42, p > .10$). Because of the interaction, in Figure 3, we shown the results of the simple main effect for the enemy condition(βenemy: $F(1, 53) = 76.26, p < .001$, βroute: $F(1, 53) = 1.87, p = ns.$) and route condition(βenemy: $F(1, 53) = 2.44, p < ns.$, βroute: $F(1, 53) = 12.17, p < .001$). In addition, because of no interaction, we shown the main effect for the enemy-route condition(conditions: $F(1, 52) = 10.87, p < .01$; variable:$F(1, 52) = 47.46, p < .001$).

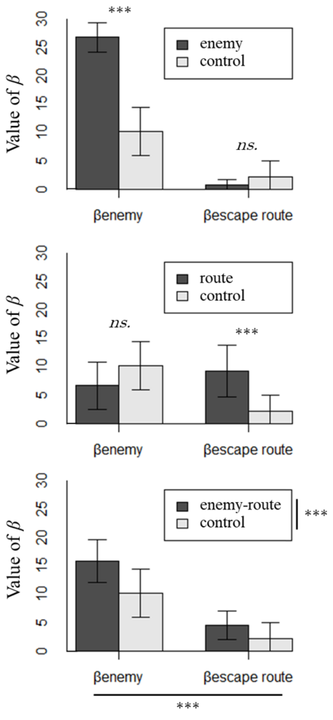

Figure 3. Comparison of β for each condition. ***: $p < .001, ns. : p > .10$

We then coded the participants according to whether they provided verbal reports about the number of enemies (e.g., “there are many ghosts around”) or escape routes (e.g., “there is only one way to go”). We created a crosstabulation table of the verbal reports with the presence of the description (2) × experimental condition (4) (TABLE I). χ^2 tests revealed significant differences between the descriptions of both the enemies ($\chi^2(3)$ = 34.10, $p < .01$) and escape routes ($\chi^2(3)$ = 26.26, $p < .01$). A residual analysis demonstrated that the ratio of descriptions of enemies was higher for the enemy and enemy-route conditions, and lower for the route condition. In contrast, the ratio of descriptions of routes was low for the enemy condition and high for the route condition; however, there was no difference for the enemy-route condition.

TABLE I
FREQUENCY AND RESIDUALS FOR THE NUMBER OF VERBAL REPORTS

A. Number of participants who described the enemy

	enemy	**route**	**enemy-route**	**control**
R	24(3.46)↑	10(−5.26)↓	24(2.34)↑	20(−0.51)
NR	0(−3.46)↓	17(5.26)↑	2(−2.34)↓	8(−0.51)

B. Number of participants who described the route

	enemy	**route**	**enemy-route**	**control**
R	1(−4.82)↑	19(3.25)↓	13(0.77)	14(−0.80)
NR	26(4.82)↓	8(−3.25)↑	13(−0.77)	14(−0.80)

[a] Direction of difference is noted for items with $p < .05$.

[b] R means Reported. NR means not Reported.

IV. DISCUSSION AND CONCLUSION

The analysis of the mental state estimation (Figure 3) demonstrated that even for the route condition, where the agent’s input represented low-saliency information, the participants were able to focus on that information and estimate the agent’s mental state. Similar results were obtained for the enemy-route condition, where the agent’s input information comprised both high- and low-saliency information. These results suggest that the participants could estimate agent’s mental states by correctly focusing on the low-saliency information regardless of the presence of high-saliency information, thereby providing an answer to RQ1.

The analysis of verbal reporting (TABLE.IB) demonstrated that low-saliency information, i.e., the route information, is less likely to be described for the enemy-route condition. This means that when agents use both high- and low-saliency information, the participants could provide a correct estimation but could not give a verbal report, thereby answering RQ2.

This study provides the following two contributions. First, we have demonstrated that estimation of the agent’s mental state can be conducted by focusing on the correct information regardless of the saliency of the input information or the combination of information. This means that participants can estimate mental states by focusing on less salient information in a simple situation where only two pieces of information are used by the agent. However, in real-world situations, users must update their estimation of the agent’s mental state according to the changing situation, e.g., during version upgrades. In addition, many practical systems use more input information. Examining the influence of the saliency of input information in such diverse situations would make it possible to elaborate on the discussion of the influence of saliency on correct estimation of mental states.

Second, we have demonstrated that the participants could estimate mental states correctly but faced difficulty reporting them verbally. This finding has implications for methods to investigate users’ evaluations and understanding of autonomous agents. Specifically, if verbal methods are utilized to survey users’ understanding, the evaluator may underestimate their understanding of autonomous agents. Thus, the impact of such verbalization when examining methods for surveying users’ understanding of autonomous agents must be considered.

ACKNOWKEDGMENT

Support for this work was given by JSPS KAKENHI Grant Number 22H03912, 22H00211 and Toyota Motor Corporation (TMC). However, note that this article solely reflects the opinions and conclusions of its authors and not TMC or any other Toyota entity.

REFERENCES

[1] N. Epley, A. Waytz, and J. T. Cacioppo, “On seeing human: A three-factor theory of anthropomorphism.” *Psychological Review*, vol. 114, pp. 864–886, 2007, doi:10.1037/0033-295X.114.4.864.

[2] C. M. Chang, K. Toda, X. Gui, S. H. Seo, and T. Igarashi, “Can eyes on a car reduce traffic accidents?” ACM, 9 2022, pp. 349–359, doi:10.1145/3543174.3546841.

[3] S. Matsubayashi, H. Terai, and K. Miwa, “Development of a driving model that understands other drivers’ characteristics.” Springer International Publishing, 2020, pp. 29–39, doi:10.1007/978-3-030-50537-0_3.

[4] C. L. Baker, R. Saxe, and J. B. Tenenbaum, “Action understanding as inverse planning,” *Cognition*, vol. 113, pp. 329–349, 12 2009, doi:10.1016/j.cognition.2009.07.005.

[5] G. Gergely and G. Csibra, “Teleological reasoning in infancy: the naive theory of rational action,” *Trends in Cognitive Sciences*, vol. 7, pp. 287–292, 7 2003, 10.1016/S1364-6613(03)00128-1.

[6] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature Reviews Neuroscience*, vol. 2, pp. 194–203, 3 2001, doi:10.1038/35058500.

[7] N. Rubin, “Figure and ground in the brain,” *Nature Neuroscience*, vol. 4, pp. 857–858, 9 2001, doi:10.1038/nn0901-857.

[8] D. D. Hoffman and M. Singh, “Salience of visual parts,” *Cognition*, vol. 63, pp. 29–78, 4 1997, doi:10.1016/S0010-0277(96)00791-3.

[9] T. C. Kershaw and S. Ohlsson, “Multiple causes of difficulty in insight: The case of the nine-dot problem.” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 30, pp. 3–13, 2004, doi:10.1037/0278-7393.30.1.3.

[10] L. J. Ball, J. E. Marsh, D. Litchfield, R. L. Cook, and N. Booth, “When distraction helps: Evidence that concurrent articulation and irrelevant speech can facilitate insight problem solving,” *Thinking & Reasoning*, vol. 21, pp. 76–96, 1 2015, doi:10.1080/13546783.2014.934399.

[11] J. W. Schooler, S. Ohlsson, and K. Brooks, “Thoughts beyond words: When language overshadows insight.” *Journal of Experimental Psychology: General*, vol. 122, pp. 166–183, 6 1993, doi:10.1037/0096-3445.122.2.166.

[12] J. R. de Leeuw, “jspsych: A javascript library for creating behavioral experiments in a web browser,” *Behavior Research Methods*, vol. 47, pp. 1–12, 3 2015, doi:10.3758/s13428-014-0458-y.