Fixed Pattern Deviation Hypothesis of Intention Attribution

Kazunori Terada[1], Seiji Yamada[2], and Akira Ito[1]

***Abstract*— A large variety of behavioral cues which invoke intention attribution to inanimate entities are reported. Although agents, including software agents and physical robots, have already possessed these features, many people still can not feel mind to them. We hypothesize that the inability of attributing intention toward agents is caused by interpreting agent's behavior as fixed pattern. In other words, a human attributes intention to an entity when it deviates fixed pattern of behavior in an efficient way (fixed pattern deviation hypothesis). In the present study we theoretically derived the fixed pattern deviation hypothesis and tested the hypothesis with human participants. We prepared an on-line experimental system in which a participant and an agent play a repeated penny-matching game with a bonus round. We then conducted experiments in which different opponent agents (human or robot) change their strategy during the game. The experimental results indicated that, as expected, adaptation is faster when a human is competing with robot than with another human. This implies that a human expect intentional deviation of fixed pattern of behavior against an human and expect deterministic and algorithmic behavior against robot.**

I. INTRODUCTION

A large variety of behavioral cues which invoke intention attribution to inanimate entities are reported, including rationality [1], [2], [3], goal or goal-directedness [4], [5], self-propelled motion [6], [7], [8], [9], [10], equifinality [1][11], spatial contingencies [12], violation of Newton's laws [13], [14], situatedness [14], [15] and motion interruption [16]. Although agents, including software agents and physical robots, have already possessed these features, many people still can not feel mind to them. We hypothesize that the inability of attributing intention toward agents is caused by interpreting agent's behavior as fixed pattern. In other words, a human attributes intention to an entity when it deviates fixed pattern of behavior in an efficient way (**fixed pattern deviation hypothesis**). Although contingent behavior, for example, may cause intention attribution toward an agent in the beginning of interaction, it will extinct when the behavior assumed to be generated by computer program. In the present study we investigated the fixed pattern deviation hypothesis with a repeated penny-matching game with a bonus round.

The fixed pattern deviation hypothesis is theoretically derived from rationality principle [1], [3]. Dennett [17] and Gergely et al. [1],[18],[19] suggest that rationality of an action plays a key roll in inferring goals of an agent. An agent initiates action on the basis of desire derived from inborn instincts (self-preservation and reproduction), and then makes plans according to beliefs about the situation. The agent will choose to perform particular instrumental action which will lead to its goal in the most rational manner, such as one that requires least effort or takes shortest path. The observer should also consider the rationality of the behavior of the agent in order to predict its future behavior, because the observer can only perceive surface behavior of the agent, and perceived behavior per se never indicate what action will be performed in the next moment.

Gergely et al. hypothesized the goal inference process of an infant as follows [1]: the infant start to monitor the agent's action, and discover an equifinal outcome. The observed equifinal outcome could provide then the specific content of the intention to be attributed to the agent. He suggest that although equifinality [11] contribute to identifying and attributing a goal to an agent, it will not allow the infant to anticipate the agent's specific future action in a new situation. This is because knowing the agent's goal will provide no information as to which of the multiple possible means actions that could lead to the goal the agent will choose to perform.

The rationality (efficiency) principle implies behavior variability principle [18]. Most efficient behavior is not most efficient in a different situation. As a result, an agent keeps to find another behavior suitable for the given situation. This process includes change in behavior. An rational agent must always evaluate the efficiency of the behavior. Then the agent must change its behavior when the situation changed and the current behavior is no longer most efficient in the changed situation. The change in behavior is observed by another agent before the behavior completed. Threfore the observed change in behavior of the target agent might trigger observer's rational inference.

The behavior variability principle is implicitly considered in the past researches. In the studies by Luo and Baillargeon [7] and Shimizu and Johnson [20], infants attributed a goal to an agent when it choose its action freely. The most effective cue for goal-directedness in Biro and Leslie's [21] study was also the variability of the goal approach. According to Csibra et. al [18] "Evidence for 'freedom' and for the capability of changing the course of action (i.e., what Premack & Premack [22] called 'motorcompetence') seems to be sufficient for infants to identify an object as an agent and to treat it as worthy of goal attribution".

Goal-directedness of behavior implies that a variety of behavior will lead to the same single outcome. Not only intentional agent but also physical phenomenon and machines have the behavioral variety. Trajectories of a falling rock, for example, differs depending on the situation. Routes of a

[1]K. Terada and A. Ito are with the Faculties of Engineering, Dept. of Information Science, Gifu University, 1-1 Yanagido, Gifu, 501-1193, Japan. Terada: terada@gifu-u.ac.jp, Ito: ai@gifu-u.ac.jp

[2]S. Yamada is with the National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda, Tokyo, Japan 101-8430

motor-driven toy car differs depending on the shape of the road surface and obstacles. However, although the outcomes of the falling rock and the moving toy car differs, only an intentional agent can pursue the same goal and reach the same single outcome [11].

A goal-directed rational agent sometimes produces fixed pattern of behavior. Fixed pattern of behavior is usually evolutionally shaped and produced by a neural network known as the innate releasing mechanism in response to an external sensory stimulus. Reinforcement learning agents also produce fixed pattern of behavior. Exploiting a reward in a most efficient way leads an optimal action sequence and the agent need not change it's behavior unless the situation is changed.

Design process is creating optimal fixed pattern of object's behavior to fulfil one's desire. Fixed pattern of object's behavior is realized by algorithm or mechanism. These underlying principle guarantees constancy of input-output relations of behaviors. Although physical phenomenon is governed by law such as gravity, physical objects never produce fixed pattern of behavior without intentional design because of chaotic properties.

From the above discussion we can make a clear distinction between intentional agents and machines in terms of behavior recognition. The distinction is whether it can deviate fixed pattern of behavior or not. Although both intentional agents and machines are able to produce fixed pattern of behavior, only agents are able to violate the fixed pattern. We hypothesize that the core concept of intentionality is the intentional deviation of fixed pattern of behavior.

II. Experiments

In this experiments we tested the fixed pattern deviation hypothesis which leads to the following hypothesis:

H A human expect intentional deviation of fixed pattern of behavior against a human and expect deterministic and algorithmic behavior against robot.

The hypothesis was tested through an experiment with human participants in a competitive game. In the game, opponent agents (human or robot) efficiently deviated fixed pattern in a critical situation of the game. The deviation was designed to be easily recognized as deception if participants of the game clearly understood the rules of the game. If the obvious deceptive behavior is treated as intentional one, a human expects further change in behavior, therefore, adaptation to the deviation is slower when a human is competing with a human. If the deceptive behavior is treated as deterministic and algorithmic one, a human expect pattern fixedness of the future behavior, therfore, adaptation to the deviation is faster when a human is competing with a robot.

A. Penny-matching game with bonus round

A repeated penny-matching game with a bonus round was created for validating the hypotheses. The penny-matching game is a zero-sum game and is played between two players, players A and B. Each player has a penny and must secretly turn the penny to heads or tails. The players then reveal their own choices simultaneously. If the pennies match (both heads or both tails) player A keeps both pennies and gets to keep player B's penny (+1 for A, -1 for B). If the pennies do not match (one heads and one tails), player B keeps both pennies (-1 for A, +1 for B). While this game has no pure strategy Nash equilibrium, the unique Nash equilibrium of this game is in mixed strategies: each player chooses heads or tails with equal probability.

TABLE I

Two strategies (straightforward and deceptive) in penny-matching game with bonus round used in experiment. Black and white circles represent heads and tails, respectively. Only difference between straightforward and deceptive strategy is choice in sixth round.

round		1	2	3	4	5	6
payoff		1	1	1	1	1	20
straightforward	uniform	●	●	●	●	●	●
	alternate	●	○	●	○	●	○
deceptive	uniform	●	●	●	●	●	○
	alternate	●	○	●	○	●	●

We modified the game rules so that players are able to use a deceptive strategy. The penny-matching game was played repeatedly. One game consists of six rounds, and ten games were played in the experiment. The payoff of the sixth round in every game is increased twenty times: the bonus round. An apparent reasonable strategy for this game is to strive to win the bonus round and abandon the other five rounds because of the large payoff gap. A player is able to trap the opponent by making a series of choices during the normal five rounds so that the opponent's prediction of the player's sixth choice will be wrong. Note that the unique Nash equilibrium of the penny-matching game with a bonus round is still a mixed (random) one even though the bonus round is added.

The agents used as opponents in our experiment used only the two strategies (straightforward and deceptive) listed in Table I. The two strategies are realized with two series of uniform and alternate choices. Uniform and alternate choices in the first five rounds were suggestive of the opponent; i.e., those exposed to the obvious trapping choices are forced to anticipate that the opponent's sixth choice will be either deceptive (violating regularity) or straightforward (keeping regularity).

B. Experimental setup and measurement

The game was implemented with JavaScript and HTML and played in a Web browser (Firefox). Figure 1 shows the game interface. A Flash video of the opponent (robot or human) is displayed at the top of the interface. The bear-like robot moves its head and arms randomly in the video. The behavior of the human was recorded when he played the game in advance. The agents' behaviors, such as choosing a side of the coin, are automatically controlled by JavaScript program. Participants were told that the opponent is online. To make the participants believe the game is online, the participants' faces captured using a web camera mounted on the monitor were displayed at the bottom of the interface.

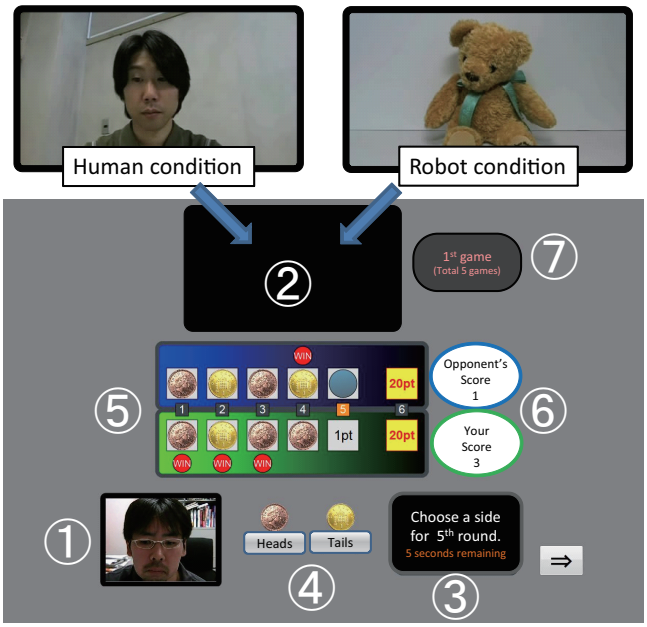

Fig. 1. Interface of on-line experimental system: 1) participant's face, 2) opponent agent's appearance, 3) remaining time, 4) choice buttons, 5) histories of both players' choices, 6) scores for both players, and 7) game number.

Participants were instructed to click the button corresponding to his/her choice within 10 seconds for every round. Scores for both players are shown in the interface. The choices of both players remain displayed so that the participant is able to recognize the opponent's strategy. Before the main experiment, all participants played five games only for training.

A single-factor two-level between-subject experimental design was used. Twenty-eight graduate and undergraduate students participated in the study. Participants were randomly assigned to either a robot or human opponent condition and were seated in front of a desktop computer. Participants were informed that the point of the experiment was to access the usability of an online game system.

A total of eight series of choices including another four series beginning from tails equivalent to the four series shown in Table I were used. The agents used a straightforward strategy, one of the four series of straightforward strategy was randomly selected, in the first three games and a deceptive one for the rest. This is the change in agent's behavior strategy during the game, and observing human adaptation to it is our aim of the experiment.

The outcomes of the sixth round for all ten games were recorded because the participants' expectation of their opponents' strategies, i.e., which strategy, straightforward or deceptive, did they expect, was of interest. Winning the sixth round against an agent indicates that the participants expectation of the agent's strategy is correct.

III. Results

The percentages of the participants who won the sixth round for all ten games are shown in Figure 2. Almost

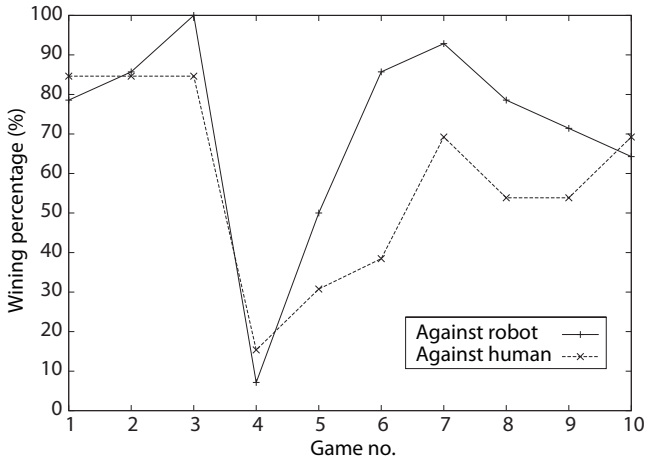

Fig. 2. Percentage of participants who won sixth round of ten games.

all the participants lost the fourth game because the agents changed their strategy from straightforward to deceptive. After the fourth game, the winning percentage with both opponent agents gradually recovered to the level of the third game. A chi-square test was conducted to investigate whether the winning percentages of the fourth to tenth games were different from that of third one (see Table II). The results indicate that at least two games were required to adapt to the change in strategy of both agents.

The differences between recovery speeds of the winning percentage between those playing against the robot and those against the human agent (plotted on the graph in Figure 2) were statistically confirmed. Table II indicates that those who played against a human agent required one more game for adapting to their opponent's change in strategy than those who played against a robot. This implies that participant's adaptation speed against a human is significantly slower than against a robot, confirming hypothesis **H**.

IV. Discussion

The participants did not equate the robot to the human in terms of adaptation speed against the opponent's change in strategy. While the strategy equally changed in the fourth game for both opponents, what caused the inequality in participant's adaptation speed is the difference in the video the participants watched. Slow adaptation of the participants who played against the human indicates that they were cautious but not optimistic against the opponent's strategies after the strategy change in the fourth game. This is because the human appearance made the participants anticipate another change in the human opponent's strategy, and the robot appearance gave the participants the impression that its strategy was stable. The behavior of the designed artifacts, including robots and computers, is governed by laws such as mechanical, algorithmic, or structural constraints; therefore, the input-output relations of the artifacts are stable. This prototypical concept of artifact's regulated behavior might make the participants expect the simpler and less complicated strategy of the robot.

Game no.	4	5	6	7	8	9	10
Against robot condition	**24.26, p<.01**	**19.33, p<.01**	2.15, n.s.	1.04, n.s.	3.36, n.s.	**4.67, p<.05**	**6.09, p<.05**
Against human condition	**14.29, p<.01**	**7.34, p<.01**	**5.60, p<.05**	0.85, n.s.	2.80, n.s.	2.80, n.s.	0.85, n.s.

TABLE II

CHI-SQUARE VALUES (DF=1) AND SIGNIFICANCE LEVELS FOR THE WINNING PERCENTAGES OF FOURTH TO TENTH GAMES COMPARED WITH THAT OF THE THIRD ONE.

There are two causes for the gradual increase in the winning percentage in the adaptation phase: 1) Each participant changed his/her strategy just once in the adaptation phase (during games 5 to 7) and the timings of the changes varied across the participants. 2) Each participant changed his/her strategy plural times to find the most suitable strategy. To investigate these causes, we compared the number of strategy changes in the adaptation phase in two conditions. The average number of strategy changes by the participants was 1.07 under an "against robot" condition and 1.57 under an "against human" condition was 1.57, so there was a marginally significant difference ($t(26) = 1.919$, $p = 0.066$). This suggests that strategy changes by the participants were one-shot deterministic under an "against robot" condition bat indecisive and exploratory under an "against human" condition.

We consider the drop of winning percentage in the eighth game could be accounted for by meta-adaptation. Our game design requires at least two levels to be adapted. The first is adaptation for regularity of the agent's choice in the first five rounds. The second is adaptation for strategy change in the fourth game. The third, which was beyond our expectations, is an adaptation for periodic change in an agent's strategy. Participants might think the agent's strategy changes once every three games and so expect to change again in the seventh game.

Not only the appearance but also the behavior of the agents differed between the two videos. We did not separate the appearance and motion factors in our experiment. Investigating which factor contributes to the change in human adaptation speed to an agent is for future work.

V. CONCLUSION

In the present study we theoretically derived the fixed pattern deviation hypothesis and tested the hypothesis with human participants. We conducted a penny-matching game with a bonus round in a web browser for investigating how humans adapt to an opponent agent's change in strategy. The experimental results indicated that, as expected, adaptation is faster when a human is competing with robot than with another human. This implies that a human expect intentional deviation of fixed pattern of behavior against an human and expect deterministic and algorithmic behavior against robot.

REFERENCES

[1] G. Gergely, Z. Nádasdy, G. Csibra, and S. Bíró, "Taking the intentional stance at 12 months of age," *Cognition*, vol. 56, no. 2, pp. 165–193, Aug 1995.

[2] K. Kamewari, M. Kato, T. Kanda, H. Ishiguro, and K. Hiraki, "Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion," *Cognitive Development*, vol. 20, no. 2, pp. 303–320, 2005.

[3] G. Csibra, "Goal attribution to inanimate agents by 6.5-month-old infants," *Cognition*, vol. 107, no. 2, pp. 705–717, 2008.

[4] W. H. Dittrich and S. E. G. Lea, "Visual perception of intentional motion," *Perception*, vol. 23, no. 3, pp. 253–268, 1994.

[5] P. D. Tremoulet and J. Feldman, "Perception of animacy from the motion of a single object," *Perception*, vol. 29, no. 8, pp. 943–951, 2000.

[6] S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind*. The MIT Press, 1995.

[7] Y. Luo and R. Baillargeon, "Can a self-propelled box have a goal ? " *Psychological Science*, vol. 16, no. 8, pp. 601–608, 2005.

[8] Y. Luo, L. Kaufman, and R. Baillargeon, "Young infants' reasoning about physical events involving inert and self-propelled objects," *Cognitive Psychology*, vol. 58, no. 4, pp. 441–486, 2009.

[9] D. Premack and A. J. Premack, "Moral belief: Form versus content," in *Mapping the mind: Domain specificity in cognition and culture*. Cambridge: Cambridge University Press, 1994, pp. 149–168.

[10] F. Heider and M. Simmel, "An experimental study of apparent behavior," *The American Journal of Psychology*, vol. 57, no. 2, pp. 243–259, 1944.

[11] F. Heider, *The Psychology of Interpersonal Relations*. Lawrence Erlbaum Associates, 1958.

[12] J. N. Bassili, "Temporal and spatial contingencies in the perception of social events," *Journal of Personality and Social Psychology*, vol. 33, no. 6, pp. 680–685, 1976.

[13] B. J. Scholl and P. D. Tremoulet, "Perceptual causality and animacy," *Trends in Cognitive Sciences*, vol. 4, no. 8, pp. 299–309, Aug 2000.

[14] R. Gelman, F. Durgin, and L. Kaufman, "Distinguishing between animates and inanimates: not by motion alone," in *Causal cognition : a multidisciplinary debate*, D. Sperber, D. Premack, and A. J. Premack, Eds. Oxford University Press, 1995, ch. 6, pp. 150–184.

[15] P. D. Tremoulet and J. Feldman, "The influence of spatial context and the role of intentionality in the interpretation of animacy from motion," *Perception & Psychophysics*, vol. 68, no. 6, pp. 1047–1058, 2006.

[16] T. Gao and B. J. Scholl, "Chasing vs. stalking: Interrupting the perception of animacy," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 3, pp. 669–684, 2011.

[17] D. C. Dennett, *The Intentional Stance*. Cambridge, Mass, Bradford Books/MIT Press, 1987.

[18] G. Csibra and G. Gergely, "'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans," *Acta Psychologica*, vol. 124, pp. 60–78, March 2007.

[19] G. Gergely, H. Bekkering, and I. Király, "Rational imitation in preverbal infants." *Nature*, vol. 415, no. 6873, p. 755, Feb 2002.

[20] Y. A. Shimizu and S. C. Johnson, "Infants' attribution of a goal to a morphologically unfamiliar agent," *Developmental Science*, vol. 7, no. 4, pp. 425–430, 2004.

[21] S. Biro and B. Hommelb, "Becoming an intentional agent: Introduction to the special issue," *Acta Psychologica*, vol. 124, pp. 1–7, January 2007.

[22] D. Premack and A. J. Premack, "Motor competence as integral to attribution of goal," *Cognition*, vol. 63, no. 2, pp. 235–242, May 1997.