Reinforcement Learning, not Supervised Learning, Can Lead to Insight

Arata Nonami (arata.nonami@gmail.com)
Department of General System Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Haruaki Fukuda (haruaki@idea.c.u-tokyo.ac.jp)
Department of General System Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Yoshiyuki Sato (yoshi.yk.sato@gmail.com)
Department of General System Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Kazuyuki Samejima (samejima@tamagawa.ac.jp)
Brain Science Institute, Tamagawa University, 6-1-1 Tamagawa Gakuen, Machida, Tokyo 194-8610, Japan

Kazuhiro Ueda (ueda@gregorio.c.u-tokyo.ac.jp)
Department of General System Studies, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan

Abstract

This study examined the differences among individuals in the performance of insight problem solving. The problem-solving characteristics of an individual seemed to be dependent on what and how they had learned. Thus, we compared the performances of insight problem solving between reinforcement and supervised learners. The results showed that the performances of reinforcement learners were better than those of supervised learners, although the non-insight problem solving performance of both learner types was comparable. This result suggests that insight might be supported by the cognitive mechanisms underlying reinforcement learning. In particular, we speculate that the degree of exploration, by which reinforcement learning is characterized, might have an impact on the performance of insight problem solving.
Keywords: insight problem solving; reinforcement learning; supervised learning; exploration

Introduction

Some people can solve daily problems insightfully while others cannot. Individual differences may impact insight displayed in daily life. If this is true, where does this difference originate?

Problem solving has been studied in cognitive science based on the framework of Newell and Simon (1972), namely, problem space theory. In their theory, problem solvers represent a problem environment as a set of possible situations to be searched to find a solution. This representation is called the problem space. The cognitive processes in many types of problem solving have been investigated based on the problem space.

Insight problem solving is characterized by a sudden solution, called the "Aha" experience after an impasse; for example, the famous anecdote on Archimedes. Kaplan and Simon (1990) suggested that an insight problem is much more difficult to be solved because its initial problem space is "ill-defined." In other words, many irrelevant or misleading features and properties are incorporated with the initial problem representation, whereas crucial aspects of the problem are omitted (Knoblich, 2009). Thus, insight problem solvers have to change their mental representation of the problem.

One dominant computational theory of insight problem solving is the representational change theory (RCT, Knoblich, Ohlsson, & Raney, 2001; Ohlsson, 1984, 1994). RCT also suggests that an insight problem solver must change the representation of the problem. RCT can explain why an impasse occurs as well as how it is broken.

Another dominant theory is criterion satisfactory progress theory (CSPT), developed by MacGregor, Ormerod, and Chronicle (2001). CSPT suggests that a balanced interplay between different kinds of heuristic values is crucial to finding the solution for an insight problem, assuming that the problem space is too large to be explored and thus difficult to find an appropriate heuristic or method.

Evidence exists for both theories. RCT and CSPT seem to direct attention to different aspects of insight problem solving. Öllinger, Jones, Faber, and Knoblich (2013) argued that CSPT focuses on the search process, whereas RCT focuses on the initial representation activated by prior knowledge. Both theories suggest that the nature of insight problem solving is related to the problem space.

Therefore, we can assume that individual differences in the performance of insight problem solving are related to individual differences in the problem space. The initial problem space, which is "ill-defined" for an insight problem, should be based on prior knowledge and experience. This approach leads to the notion that individual differences in insight could depend on how learners learned in the past.

In computer science, there are three main classes of learning algorithms: supervised, unsupervised, and reinforcement learning. As unsupervised learning does not have a goal, we ignore it here. Reinforcement learning is characterized as learning by trial and error, whereas supervised learning is based on exemplars. When supervised

learning and reinforcement learning are expected to show equal performances, which learning style is employed for the exploration of a problem space? An individual's preferred or familiar style might be ideal. This preferred learning style can be assumed to affect the structure of the problem space and how to search there.

Based on this idea, we investigated the relationship between an individual's performance of insight problem solving and preferred learning style. The findings shed light on the cognitive processing recruited for insight. This study could thus be a bridge between insight studies in cognitive science and learning studies in computer science.

Purpose of the Study

To investigate whether differences in insight among individuals could depend on their preferred learning style, we conducted two experimental tasks.

The first was an insight problem solving task that required participants to change the representation of a given problem. The second was a simple learning task where reinforcement and supervised learning were likely to be equally effective. The participants were classified into two groups, namely, reinforcement learners (RLs) and supervised learners (SLs), based on their results in the learning task. We then compared the insight problem solving performance between the two groups.

Method

Participants

Forty-five undergraduate students (36 females and 9 males, 19.98 ± 0.723 years old) at Aoyama Gakuin University participated in the experiments. All were unaware of the purpose of the experiments, which were conducted as approved by the Ethics Review Committee on Experimental Research with Human Subjects at the University of Tokyo's Graduate School of Arts and Sciences.

Experimental Tasks

Insight Problem Solving Task

The participants engaged in matchstick arithmetic problems, including the so-called insight problems (Komazaki & Kusumi, 2001; Knoblich, Ohlsson, Haider, & Rhenius, 1999). These problems required solvers to change their representations to arrive at a solution.

In the matchstick arithmetic problems, the participants were shown false arithmetic statements written with Roman numerals (I, II, III, etc.), arithmetic operators (+, -), and equal signs (=), which consisted of matchsticks.

The participants were asked to move only one matchstick to transform the given false statement into a true arithmetic one (Figure 1).

We defined the insight problems in this study as those shown in Figure 1b. Here, the solution is to make the second "=" sign by moving the vertical stick in "+," and to create tautological equations (III=III=III). This type of problem should be "ill-defined," because an assumed initial representation seems that "an equation has only one equal sign," which did not include the path to solution. Therefore, this could considered an insight problem, which followed the definition of Kaplan and Simon (1990). We also used a "tautological equation problem" as an insight problem in this study. Non-insight problems (Figure 1a) would be solved without such difficulties caused by the change of mental representation for an initial problem space.

Learning Task

The second task for the participants was a simple learning task. This task was the simplest version of a binary choice task, which is often used in machine learning and in the field of neuroscience. During this task, the participants were forced to make a series of choices between two rewards, each of which was given stochastically and asked to maximize their accumulated outcome. Thus, the participants must learn each reward probability from its past reward history to maximize their outcome (Figure 2). The participants were instructed that each reward probability was constant in the experiment and associated with the color of options, red or green.

This learning task can be understood from two perspectives. First, this is a task in which the participants learn to make better choices from trial-and-error, by selecting options and receiving feedback in the form of rewards. From this perspective, the learning model for the task is based on reinforcement learning; the task is regarded as a kind of bandit task, which is a typical reinforcement learning problem (Sutton & Barto, 1998). Second, this is a task in which participants classify the feature of each option (red or green) into "good" or "bad" based on the success or failure of the former trials. From this perspective, the task is regarded as a concept learning task, in which supervised learning would work (Valiant, 2013).

We designed the learning task, allowing the participants to employ either of the two learning styles: reinforcement and supervised learning. Based on each participant's selection history, we estimated which of the learning styles they preferred, using computer modeling and model comparison.

Experimental Procedure

All the participants participated in both the learning and insight problem solving tasks.

First, they performed 30 trials of the learning task after a practice session. This practice session comprised 10 trials, where reward expectation was the same between both options. In the experimental session, the reward expectations for two options were 70% and 40%, respectively, which were assigned randomly to either of the two colors (red or green). Beforehand, the participants were instructed that each option had a constant reward probability throughout the experiment and that they could get a constant outcome, 10 points if rewarded. The participants selected the right or left option by pressing a button, and then feedback (a reward or no reward) was given to them in each trial.

Then they participated in the matchstick arithmetic task, which comprised 12 problems (including three insight problems). Each problem was shown to participants for 30 seconds, then the next problem was displayed automatically. Participants were asked to solve each displayed problem within 30 seconds. The order of problems was randomized, and the display was controlled by a computer.

Learning Models

To classify the participants as reinforcement learners (RLs) or supervised learners (SLs), we fitted each participant's choice history to two models that were explained in this section.

Reinforcement Learner

Reinforcement learning in computer science is defined as a dynamic algorithm that learns by interacting with its environment. The agent receives rewards and updates its expectation or value by rewards, which were better than expected, and by penalties, which were worse than expected, according to value function.

Value function ($Q(a)$) is a function that returns the value of an action when the action is input. The function continues to update itself during learning by using the difference between estimated and actual rewards.

Value (Q) for the action choosing red or green option was calculated in the reinforcement learning model as:

$$Q(a_t) \leftarrow Q(a_t) + \alpha(R_{t+1} - Q(a_t))$$

t: trial number, a : {choose red, choose green}, R_t : the magnitude of reward

Choice probability (*P*) of each action was estimated by the following softmax function:

$$P(a) = 1/(1 + \exp(-\beta(Q(a) - Q(\bar{a}))))$$

A parameter "α" is the learning rate. It is a step-size parameter of a positive fraction. It is used to progressively approximate the optimal policy. The temperature parameter "β" shows how sensitive an agent is to the difference between the values for actions.

Supervised Learner

Supervised learning in computer science is defined as the learning in which a function is inferred from labeled training data. A supervised learning model analyzes the training data and produces an inferred function. Thus, the training phase and the subsequent test phase are independent and separated obviously.

For our supervised learner, first *n* trials were determined as the test phase. In machine learning, it is difficult to determine the appropriate duration of training. However, our purpose was only to estimate the duration posteriori, thus we estimated "*n*" directly as a free parameter. We set the choice probability in the training phase to be 1/2. After the training phase, hit probabilities (*HP*) for red and green options were calculated. In the test phase, choice probability (*P*) was calculated from these hit probabilities (*HP*) by the following softmax function:

$$P(a) = 1/(1 + \exp(-\beta(HP(a) - HP(\bar{a}))))$$

a : {choose red, choose green}

We used these two models to fit each participant's data. These models are almost the simplest form in the both types of learning, to elicit characteristics of the participants' learning styles.

Data Analyses

Our interest is to test whether the performance of insight problem solving is different between RLs and SLs. For this purpose, at first, we compared the correct response rate for insight problems with that for non-insight problems, to check whether the former were more difficult to be solved the latter. Then we classified our participants as RLs or SLs according to the learning models. Finally, we compared the performance of insight problem solving between the two learning styles.

Results

Performance in Matchstick Arithmetic Problems

In general, insight problems are more difficult than non-insight ones because of the "ill-defined" problem space for the former problems. To check whether our insight problems were more difficult than the non-insight ones, we analyzed the performance for each type of problem. As a result, the correct response rate for the insight problems was lower than that for the non-insight problems ($t(44) = 8.637$, $p < 0.001$, as shown in Figure 3), as we expected. This could be attributed to our insight problems requiring problem solvers to change the mental representation of the initial problem space. This suggested that insight problems could be differentiated from non-insight ones in our task.

Classification of Participants: Reinforcement or Supervised Learner

Because the learning task can be solved both by reinforcement and supervised learning, we applied the two learning algorithms to the data of the participants' choices and compared the goodness of fit between the algorithms for each participant. Then we classified each participant either as an RL or SL, according to Akaike's information criterion (AIC; Akaike, 1973).

As a result, we classified 23 participants as RLs and 22 as SLs (Figure 4). We used this classification in our later analysis.

Moreover, we compared the learning task performances between RLs and SLs to confirm whether both learning styles would equally work well. The results showed that the RLs and SLs showed comparable scores for the learning task,

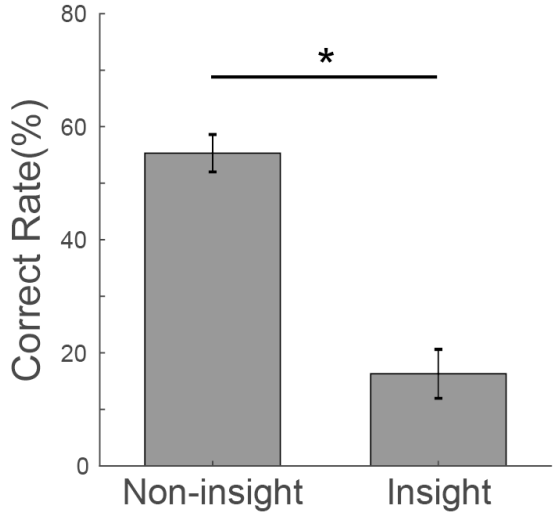

Figure 3: Correct response rates for insight and non-insight problems. Insight problems were much more difficult than non-insight ones ($t(44) = 8.637$, $p < 0.001$). Error bars indicate the standard error of the mean (SEM).

($t(43) = 1.711$, $p = 0.248$). Thus, we could say that both learning styles were beneficial to our learning task (Figure 5).

Insight Problem Solving Performance of Reinforcement and Supervised Learners

The results for the learning task showed that both types of learning existed: reinforcement and supervised learning. We investigated whether there was a difference in the performances of insight problem solving between RLs and SLs (Figure 6). For the insight problems, the correct response rate for RLs was significantly higher than that for SLs ($t(43) = 2.650$, $p = 0.011$), whereas, for the non-insight ones, no difference was observed ($t(43) = 0.517$, $p = 0.608$).

These results showed that the RLs were superior to SLs only in insight problem solving. Therefore, the nature of reinforcement learning, and not of supervised learning, has an impact on insight problem solving.

Discussion

The results showed that the RLs showed better performance than the SLs only in insight problem solving. This suggests that the bias for selecting a learning style has an influence on the results of insight problem solving.

This could not result from the difference of the participants' general abilities, because we did not find a difference in the performance of both non-insight problem solving and learning task. The nature of reinforcement learning, and not of supervised learning, might have certain advantages in solving insight problem in which problem space is "ill-defined."

Although reinforcement learning should be active in the environment, supervised learning learns from the given data.

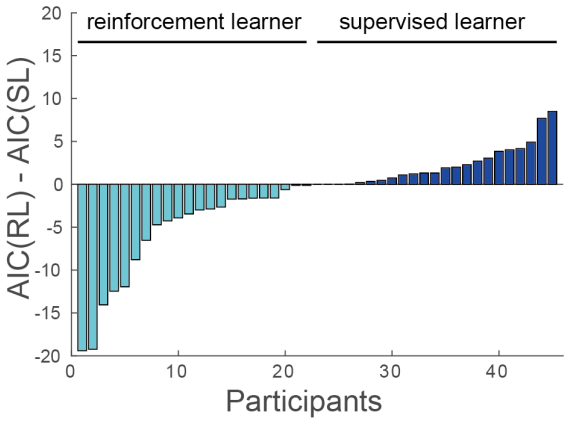

Figure 4: Difference in the value of Akaike's information criterion (AIC) between fitting by reinforcement learning model and that by supervised learning model. A value greater than zero means that the participants' behavioral data were better fit by the SL model than the RL model. Twenty-three participants were RLs, whereas 22 were SLs. Participants whose AIC difference was close to zero showed that both learning models were comparable in data fitting; when the AIC difference became larger, one model was superior to the other in data fitting.

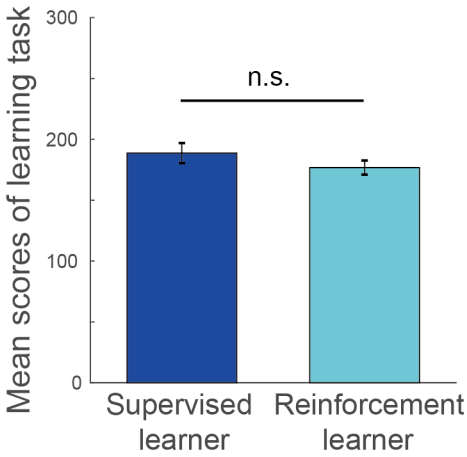

Figure 5: Mean scores of the learning task for RLs and SLs. There is no difference in the average score. Therefore, learning style is neither superior nor inferior to the other in the learning task ($t(43) = 1.711$, $p = 0.248$). The error bars indicate the SEM.

Additionally, although people can employ both learning styles, a person primarily employs one style, which seems to be related to insight problem.

The learning style which supervised learners employed for the learning task is also interpreted as "explore-then-exploit strategy" in the computer science domain (Kaelbling,

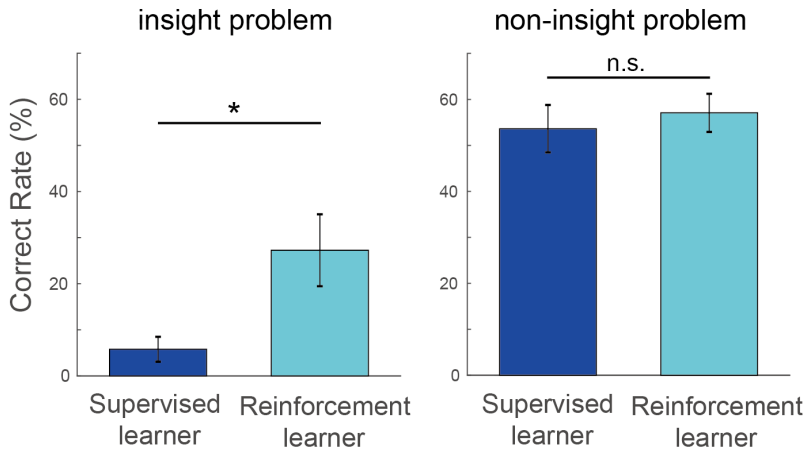

Figure 6: Correct response rates for the insight and the non-insight problems by the two types of learning style (supervised and reinforcement learners). Although the correct response rates for the non-insight problems were comparable between the SLs and the RLs ($t(43) = 0.517$, $p = 0.608$), the RLs showed better performance on the insight problems than the SLs ($t(43) = 2.650$, $p = 0.011$); the participants who preferred reinforcement learning solved the insight problems better. The error bars indicate SEM.

Littman, & Moore, 1996). Explore-then-exploit strategy is among the strategies for bandit task, which divides clearly the task into two phases, namely, exploration and exploitation phases. During exploration phase, supervised learner estimates probability of reward by random policy. Then during the exploitation phase, supervised learner fixes policy with taking optimal action. On the other hand, reinforcement learner continuously explores the environment with Boltzmann distribution even after getting sufficient number of rewards to take optimal action. The difference between two learning styles corresponds to that in exploration strategies. Therefore, the individual difference of exploration strategy can be measured by model fitness of the AIC difference in this study. As suggested by RCT and CSPT, a key in insight problem solving is searching or exploring the problem space. Taken together, we can assume that the nature of exploration in reinforcement learning could lead to the solution of an insight problem. Kaplan and Simon (1990) stated that flexibility or the willingness to try a variety of things might facilitate insight.

There are two possible explanations for the way exploration affected insight problem solving. One is related to RCT. In RCT, a solver searches the current problem space quickly while simultaneously searching an appropriate space to find a path to the solution in a meta-space, which comprises possible problem spaces. This style requires quick exploration. Because an RL becomes familiar with the solution through reinforcement learning, the structure of her/his problem space might make an extensive exploration feasible. As a result, such explorations might enable an RL to change the mental representation of a problem and obtain an insightful solution rather quickly.

Another explanation is related to CSPT. In CSPT, solvers manage different kinds of heuristic (maximization and progress-monitoring heuristics (MacGregor, Ormerod, & Chronicle, 2001) to explore a large problem space. This is a merit of reinforcement learning, which allows the reinforcement learning algorithm to maximize the reward expectation in balancing exploration for the future outcome and exploitation of the current knowledge.

In summary, we found that participants who preferred reinforcement learning showed better performance in insight problem solving. This suggested that the nature of exploration in reinforcement learning might facilitate the search for the goal in problem space. Insight has been distinguished from an incremental learning process, such as reinforcement learning, because it is characterized by a sudden solution with an "aha" experience. Our findings imply that insight and reinforcement learning might have a link in a cognitive substrate, intermediated by exploration.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number JP16H01725.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information Theory*. 267–281.

Kaelbling, L. P., & Littman, M. L. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, *22*, 374–419.

Komazaki, H., & Kusumi, T. (2002). Representation transformation and constraint relaxation in insight problem solving: Investigation on repeated matchstick algebra problem. *Cognitive Studies*, *9(2)*, 274–284.

Knoblich, G. (2009). Psychological research on insight problem solving. In *Recasting reality* (pp. 275–300). Springer, Berlin, Heidelberg.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25(6)*, 1534–1555.

Knoblich, G., Ohlsson, S., & Raney, G. E. (2001). An eye movement study of insight problem solving. *Memory & Cognition*, *29(7)*, 1000–1009.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2001). Information processing and insight: A process model of performance on the nine-dot and related problems. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27(1)*, 176–201ss.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Robertson, S. I. (2016). *Problem solving: Perspectives from cognition and neuroscience*. Psychology Press.

Ohlsson, S. (1984). Restructuring revisited: I. Summary and critique of the Gestalt theory of problem solving. *Scandinavian Journal of Psychology*, *25(1)*, 65–78.

Ohlsson, S. (1992). Information-processing explanations of insight and related phenomena. *Advances in the Psychology of Thinking*, *1*, 1–44.

Öllinger, M., Jones, G., Faber, A. H., & Knoblich, G. (2013). Cognitive mechanisms of insight: The role of heuristics and representational change in solving the eight-coin problem. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39(3)*, 931–939.

Sutton, R. S., & Barto, A. G. (1998). *Introduction to reinforcement learning*. Cambridge: MIT press.

Valiant, L. (2013). *Probably Approximately Correct: NatureÕs Algorithms for Learning and Prospering in a Complex World*. Basic Books (AZ).