Proposed Decision-Making System Based on Consciousness in Multiple Rewards and Penalties Environments

Kazuteru Miyazaki

***Abstract*—What is consciousness? Many definitions have been proposed to describe this concept. Our research aims to achieve consciousness in computers and, in particular, focuses on a decision-making system based on consciousness. After the requirements are defined, a system that satisfies these conditions is proposed. Especially, this work proposes a new decision-making system based on consciousness. This model consists of emotional memories, declarative memories, and an amygdala-like mediation component incorporating three types of modules such as agilor, contemplator, and mediator. The effectiveness of the proposed system is validated through numerical experiments.**

***Index Terms*—Consciousness system, exploitation-oriented learning (XoL), reinforcement learning.**

I. Introduction

What is consciousness? Many definitions [1] have been proposed to describe this concept. Also Marvin Minsky says that “consciousness is one of suitcase-like word that we use for many types of processes and for different kinds of purpose [2].” The study of consciousness has two aspects: understanding human brain function [3]-[7] and implementing on a computer [8]-[12]. Although the main focus is on the second aspect from the viewpoint of engineering, a successful implementation must make the most of the knowledge on brain function. On the other hand, the realization of machine consciousness may also contribute to the elucidation of consciousness in the brain.

Achieving machine consciousness presents various hurdles. A decision-making method based on consciousness has been addressed previously [10], [11]. A primary system involving *exploitation-oriented learning* (XoL) [13] or *reinforcement learning* (RL) [14] that exhibits a simple interaction with an environment was combined with a secondary system that can observe the primary system and perform any operation of this primary system, resulting in an effective interaction between these systems [11].

Daniel Kahneman, the Nobel Prize laureate in Economic Sciences, in his book, divides decision making in humans into automatic and controlled operations [15]. Fast thinking is immediate, automatic, and intuitive whereas slow thinking is deliberate, effortful, and controlled. Human memory is also categorized into emotional and declarative memories from the perspective of brain sciences [16]. While the declarative memory plays an important role in the complex but peaceful social and linguistic environment, the emotional memory reacts quickly to environmental changes during emergency situations [16]. By analogy, emotional and declarative memories correspond to automatic and controlled processes, respectively.

Amari [16] and others [17], [18] have also claimed that “switching between the two memories might occur in the amygdale”, suggesting the importance of the amygdala in human decision making.

This work considers a new *decision-making system based on consciousness*. This model consists of emotional memories (automatic processes), declarative memories (controlled processes), and an amygdala-like mediation component.

Researchers [19], [20] have proposed *MarcoPolo* as a decision-making system. MarcoPolo switches between the XoL *profit sharing* [21], [22] method and *k-certainty exploration method* [23] aimed at identifying an environment through a mediator that is designed for domain-specific reinforcement learning tasks under Markov decision processes (MDPs). However, MarcoPolo is less versatile because of its MDP specialization. It is also assumed that a kind of reward for learning is one. In this work, the author proposes a decision-making system based on consciousness in multiple reward and penalty environments and numerical experiments are conducted to assess its effectiveness.

II. The Domain

Consider an agent in an unknown environment from which an input is called *a state.* After perceiving this state, the agent selects and executes an action. The environment may also contain an internal state that is generated by the agent itself. Time is discretized in one input-action cycle. A pair of a state and an action selected in the state is called a rule. A function that maps states to actions is called *a policy.*

The problem in a decision-making system based on consciousness involves determining the action output to be selected for each sensory input. A reward or a penalty is assumed to act as a teacher to solve the problem. Rewards and penalties are received from the environment on the basis of a series of actions. A reward is given to a state or action causing transition to a state in which a specific goal is achieved. In contrast, a penalty is given when a state or corresponding action does not achieve this goal. The purpose of learning is to

Manuscript received July 10, 2014; revised October 8, 2014. This work was supported by JSPS KAKENHI Grant Number 26330267.

Kazuteru Miyazaki is with the National Institution for Academic Degrees and University Evaluation, 1-29-1 Gakuen-nishimachi Kodaira-shi, Tokyo 187-8587 Japan (e-mail: teru@niad.ac.jp).

DOI: 10.7763/IJMLC.2015.V5.494

generate a policy that continuously produces rewards without any penalty.

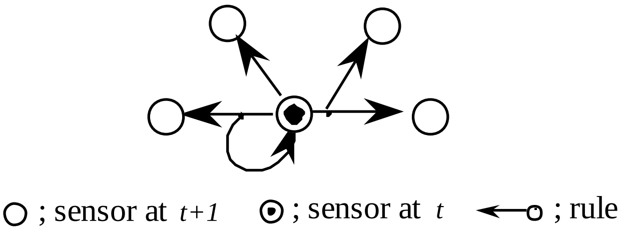

Fig. 1. Schematic representation of an environment.

The environment is treated as a collection of stochastic processes, where a sensory input corresponds to a certain state and an action corresponds to a state transition operator. Fig. 1 shows the state transition diagram representing an environment. The node with a dot denotes a sensory input at time *t*. This sensory input gives rise to three rules. Because the state transition is not deterministic, selecting the same rules does not always lead to the same state, as indicated by the branching arcs.

The learning agent does not have a priori knowledge about a complete state transition and thus, needs to learn the policy by interacting with the environment. This "goal-directed learning from interaction" is treated by RL [14] or XoL [13] methods, which may guarantee that the acquisition of the policy yield continuous rewards without any penalty, that is called a *rational penalty avoiding policy*, when the environmental class is correctly assumed.

III. Overview of the Proposed Decision-Making System

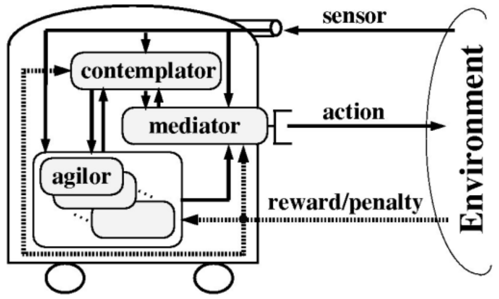

Fig. 2. Proposed decision-making process.

Fig. 2 shows the overall structure of the proposed decision-making system. Its components, or *modules,* consist of several *agilors*, one *contemplator,* and one *mediator* that decides and outputs an action. In MarcoPolo, the agilor and contemplator correspond to profit sharing and k-certainty exploration methods, respectively.

A. Agilor

An agilor is defined as a module that learns quickly. The *primary system has* previously been used as an agilor [11]. This system executes an action in response to an input from the environment. Therefore, both RL and XoL correspond to primary systems because they aim to learn a policy that determines the association between an agent and environment.

In general, the *emotional system,* which is believed to stem from the human instinct, and reflexive action represented by conditioned reflex can also be regarded as agilors. In this work, they are described as independent agilors that correspond to emotional memory [16] or automatic process [15].

Once the relationship between modules is defined, the system can be configured similar to the *subsumption architecture* [24], which typically *a*ssumes a hierarchy between modules. In this architecture, modules become more abstract when they access an upper layer. Moreover, the purpose of each layer needs to account for the purpose of the lower layer. However, in this work, the system does not adopt this hierarchical framework because each agilor is regarded as an independent module. Conflict resolution between agilors is accomplished through contemplator or mediator tasks.

B. Contemplator

While agilors represent different types of modules, such as emotion, reflection, and learning, contemplator oversees agilors or executes sophisticated calculations. The role therefore is to ensure proper action selection.

Contemplator corresponds to the working memory that is considered to exist in the human neocortex. Poorer grade animals likely execute decision making by a responsive process only because they have less or no neocortex. In contrast, humans achieve a higher level of decision making by comparing agilor contents using their large neocortical region. Contemplator may also perform a calculation of their own and, therefore correspond to declarative memory [16] or controlled processes [15].

C. Mediator

An action that can be output to the environment generally represents a mediator while agilor and contemplator correspond to candidates for each action. This module determines which action should be outputted to the environment.

The *secondary system* has previously been defined as a mediator [11]. It interacts directly with a primary system but not with an environment. In that case [11], the secondary system is generated when some unexpected situation occurs after the learning of the primary system is stable. Consequently, the agent can adapt to unforeseen circumstances without resetting everything that the primary system has acquired.

This is believed to be effective when the primary system only consist of one type of agilor, as suggested by numerical experiments [11]. However, this configuration is irrelevant when several agilors are present, such as in this work.

Here, the mediator should focus on which module should be selected because the details of decision making are completed by the agilor and contemplator modules. To this end, the mediator requires inputs from the contemplator and individual agilors along with the input-output relationship in the environment. These pieces of information provide the mediator with the ability to determine which action should be output to an environment. Remark that there is the input-output relationship with the environment in this work, whereas there is no direct interaction with the environment in the secondary system [11].

IV. Model Construction Process

A. Agilor

The agilor mainly encompasses emotion, reflection, conditioned reflex, and learning.

Emotion and reflection are inherent to living organisms. They form the basis of life that has been processed by the brain stem, such as escape from fear and homeostasis to maintain life. Therefore, they are incorporated in advance according to certain tasks.

On the contrary, the conditioned reflex is considered as a response acquired by applying the learning function to behavior based on emotion. In other words, it is equivalent to learning later in life what to avoid.

Learning a policy may also be used as an agilor function, such as learning by XoL or RL. Learning results are significantly influenced by reward or penalty settings.

B. Contemplator

The contemplator is responsible for high-cost tasks that cannot be processed by the agilor, such as search, exploration, comparison between considerable amounts of data, and computationally demanding learning tasks. The contemplator configuration rests on careful examination of individual target problems because of its problem dependence.

For example, a comparison between the contents of each module requires an identification of the modules that can be executed within a certainty factor. This may be achieved by generating a model of the environment and performing a simulation in this model. However, the generation and simulation of a model generally demand significant computational resources, resulting in a slow response. These tasks are therefore suitable for a contemplator.

Computationally costly tasks, such as long-step learning and combinational search problems, may also be assigned to the contemplator.

C. Mediator

The mediator makes a final decision on the action to be output to the environment and may be considered equivalent to the amygdala in humans. Although its implementation depends on the problem, the following configuration may be envisaged as an example.

Consider the case where the mediator makes a decision based on the signal strength of each module. The signal strength corresponds to its confidence factor. In this case, the reaching speed of the signal, which is related to the time interval at which the same module produces an output, is also important. The mediator determines which module should be selected according to speed and confidence. Therefore, parameters should be tuned depending on which of the speed or confidence is updated by learning.

Each module is assumed to provide information on the expected transition destination. The mediator adjusts the signal speed and confidence parameters by learning using the difference between the expected transition destination and actual distribution information, as well as reward/penalty signals emitted by the environment. XoL or RL may be utilized for learning.

In the next section, the effectiveness of the proposed decision-making system is verified by numerical experiments using a simplified mediator.

V. Numerical Experiments

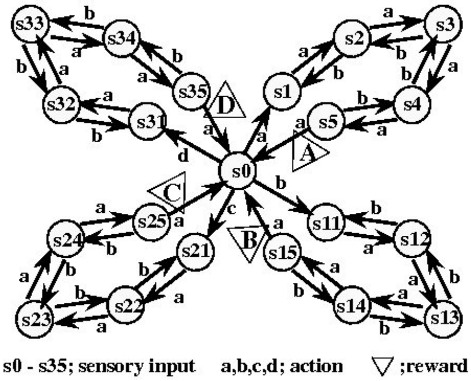

Fig. 3. Environment used in the numerical experiments.

A. How to Validate the Effectiveness of the Proposed Method

As shown in Fig. 2, the proposed decision-making system based on consciousness comprises agilor, contemplator, and mediator modules. The performance of the system is first evaluated when only one module is used before combining all three components, resulting in the following three experiments:

- Agilor only
- Contemplator only
- Proposed system

Although the contemplator may also be combined with the agilor, this configuration is excluded because an action cannot be selected from several candidates when there is no mediator.

Fig. 3 is a state transition diagram of the environment used in the experiments reflecting state transitions evaluated after the system has conducted a sufficient search of the environment. The learning agent exhibits four actions (a, b, c, d) and their four corresponding desire levels (A, B, C, D). When an action is selected, the desire levels decrease. This decrease is predetermined for each desired level. The agent can properly observe all desire levels but the function that defines the decrement of each desire level is not known in advance.

The agent receives 21 types of sensory inputs labeled s0–s35. State transitions that do not change the sensory input, such as returning to s1 after selecting action c in s1, are omitted in Fig. 3.

When the agent selects action a in s1, its transits to s3 by a probability p of 0.5 and then to new state s6 located between s1 and s2 by another probability (p = 0.5, data not shown). The same uncertainty occurs in s11, s21, and s31.

On reaching a goal, represented by triangle A, B, C, or D in Fig. 3, the corresponding desire level is increased. This increment, which is predetermined for each desire level, corresponds to a reward. On the other hand, if one level decreases to zero, all desired levels are initialized and the agent returns to s0, which represents a penalty.

In this situation, consider the problem which action should be selected in s0. The solution depends on the change in desire level resulting from selecting an action and acquiring a reward. The agent's purpose is to learn a rational penalty avoiding policy.

B. Specific Module Design

1) Agilor: A *minimum priority strategy* (MPS) [10], in which the agent always aims to achieve the minimum desire level, is adopted for agilors.

MPS is expected to be an effective strategy in many cases. However, when the decrement in desire level A is much larger than that in B, action A should be selected over B if its desire level exceeds that of B. The threshold changes according to the following factors: the decrement of desire level A, decrement of desire level B, difference between these levels, and increment of the desired level when the agent reaches the goal. Configurations that are inappropriate for MPS have been previously investigated [10] and are accounted for in this work.

2) Contemplator: The contemplator is designed using the learning based on an *avoidance list* proposed previously [25].

An avoidance list, corresponding to the set of lower layers that should not be selected in each state, is therefore established for more emphasis on penalty-based learning. This list is used to select an appropriate lower layer. When the agent receives a penalty, the following information in the previous decision-making is registered in the avoidance list.

- Type of goal that needs to be achieved.
- Desire level value of goal that needs to be achieved.
- Desire level value of goal that received the penalty.

This registration should take into account the interaction between "the goal that needs to be achieved" and "the goal that received a penalty." For example, when the agent that aims to satisfy the desire level A receives a penalty about desire level B and desire level A is included in the penalty, the avoidance list is updated with the information in the most recent decision-making point that desired level A had not been aimed. The avoidance list is also updated when its range is broadened.

3) Mediator: MPS is expected to be effective for many cases in this problem. Therefore, the mediator selects the agilor output as long as there are no problems.

However, if the agent receives a penalty when it follows an action from the agilor in an environment, the mediator selects the contemplator output in the same environment.

Although a similar switching process has been previously followed [25], it is not possible to switch between agilor and contemplator for each type of environment since a separate module such as a mediator does not exist.

C. Experimental Setup

In the numerical experiments, the lower-level learning that aims to achieve each goal from each state (s0, s1,..., s35) through the lowest number of actions is completed.

Initial desire levels amount to 100. During these experiments, "the decrement in each desire level after the agent outputs an action" and "the number of rewards acquired when the agent achieves a goal," which corresponds to "the increment in desire level related to this achievement," are changed. Environments 0, 1, and 2 shown in Table I are repeated each time the number of rewards and penalties reaches 300,000. Previous knowledge has shown that MPS is effective for environments 1 and 2 but cannot prevent a penalty in environment 0 [10]. All desired levels are initialized and the agent transits to s0 when the environment has changed.

TABLE I: VARIATION OF ENVIRONMENTS

	Decrement in the desired levels				Increment in the desired levels			
	A	B	C	D	A	B	C	D
Environment 0	1	1	1	5	60	60	60	60
Environment 1	1	1	1	1	80	80	80	80
Environment 2	1	1	3	3	80	60	60	60

Each experiment proceeds 15 $\times 10^7$ times until an action is selected. These experiments are performed 100 times using different random seeds and the average number of rewards and penalties is evaluated.

D. Results and Discussion

When only the agilor is used, unlike environments 1 and 2, environment 0 cannot completely avoid a penalty, in agreement with a previous study [10].

Experimental results involving the proposed method and contemplator only are shown in from Fig. 4 to Fig. 9, respectively. Vertical axes are not aligned for visibility.

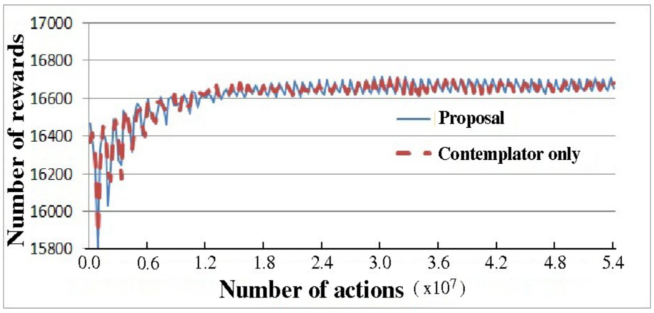

Fig. 4. Number of rewards in environment 0.

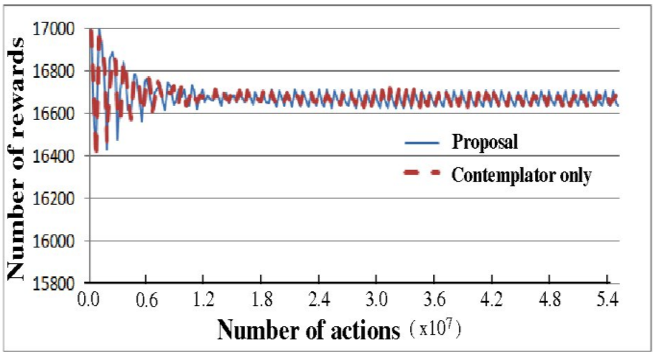

Fig. 5. Number of rewards in environment 1.

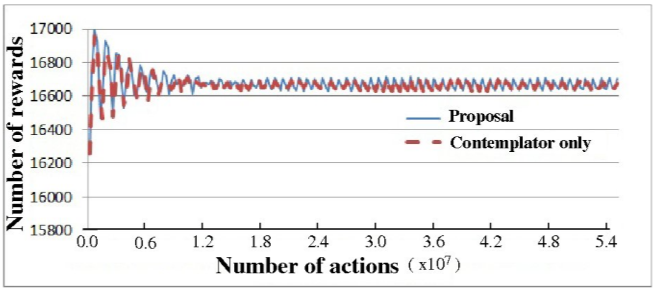

Fig. 6. Number of rewards in environment 2.

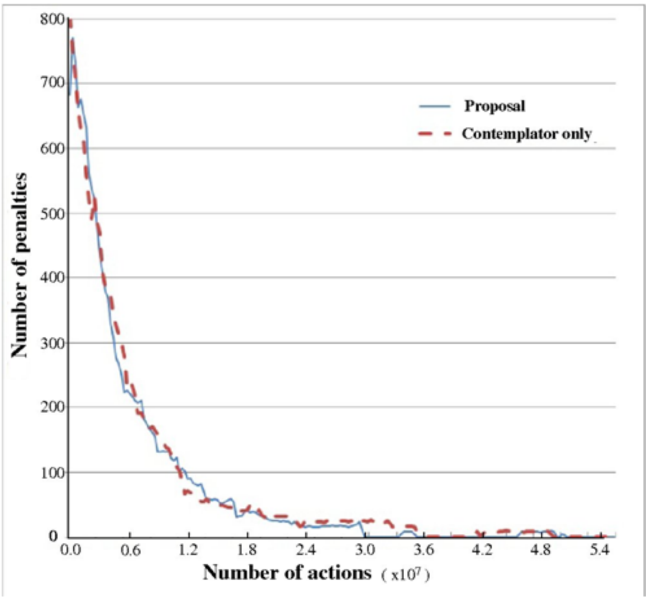

Fig. 7. Number of penalties in environment 0.

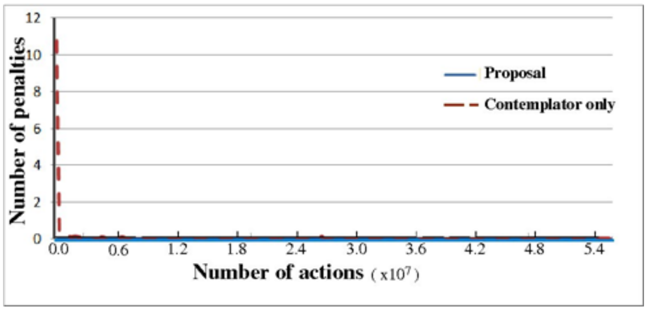

Fig. 8. Number of penalties in environment 1.

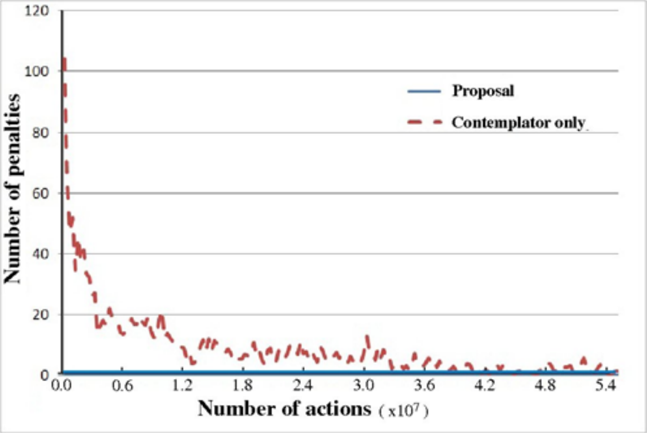

Fig. 9. Number of penalties in environment 2.

Fig. 4, 5 and 6 show that there is little difference between the proposed method and contemplator only for acquiring rewards. Moreover, little difference is observed between these methods in environment 0 as shown in Fig.7. In experiments involving only the contemplator, the agent receives a penalty in the MPS-treatable environments 1 and 2 (Fig. 8 and 9). In contrast, the proposed method avoids penalties in all environments.

When only the contemplator is used, many actions are required to receive a penalty because the contemplator aims to learn from an avoidance list drawn for a simple MPS-treatable environment. However, in the proposed method, the agent can select an appropriate module on the basis of the difficulty of the environment because it can switch from the avoidance list to MPS using the mediator. These results suggest the effectiveness of the proposed system comprising agilor, contemplator, and mediator modules.

VI. Conclusion

This work proposed a decision-making system based on consciousness incorporating three types of modules such as agilor, contemplator, and mediator. The effectiveness of the proposed system was confirmed by numerical experiments.

In future work, this system will be applied to real-world problems, such as keepaway tasks [26], [27], the biped walking robot [28], and the course classification support system at National Institution for Academic Degrees and University Evaluation [29].

Acknowledgment

The author wishes to thank Prof. Junichi Takeno for his helpful comments about consciousness systems.

References

[1] S. Blackmore, *Conversations on Consciousness: What the Best Minds Think About the Brain, Free Will, and What It Means to Be Human*, Oxford Univ Pr, 2007.

[2] M. Minsky, *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*, Simon & Schuster, 2006, p. 95.

[3] C. Koch, *The Quest for Consciousness: A Neurobiological Approach*, Roberts & Co, 2004.

[4] L. Nick, *Life Ascending: The Ten Great Inventions of Evolution*, W W Norton & Co Inc., 2010, pp. 232-259.

[5] S. Blackmore, *Consciousness: An Introduction*, Oxford Univ Pr, 2003.

[6] W. Penfield, *Mystery of the Mind: A Critical Study of Consciousness and the Human Brain*, Princeton Univ Pr, 1978.

[7] B. Libet, *Mind Time: The Temporal Factor in Consciousness*, Harvard University Press, 2005.

[8] J. Hawkins and S. Blakeslee, *On Intelligence*, Griffin, 2005.

[9] M. Kawato, *From "Understanding the Brain by Creating the Brain" toward Manipulative Neuroscience*, Philosophical Transactions of the Royal Society B, 2007.

[10] K. Miyazaki, "Research of a decision making method based on consciousness in multiple rewards environments," in *Proc. 39th SICE Symposium on Intelligent Systems*, 2012, pp. 95-98.

[11] K. Miyazaki and J. Takeno, "A study on the necessity of a secondary system in the consciousness system," in *Proc. 2014 Annual International Conference on Biologically Inspired Cognitive Architectures*, to be published.

[12] J. Takeno, *Creation of a Conscious Robot: Mirror Image Cognition and Self-Awareness*, Pan Stanford Publishing, 2012.

[13] K. Miyazaki and S. Kobayashi, "Exploitation-oriented learning PS-r#," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 6, pp. 624-630, November 2009.

[14] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book, MIT Press, 1998.

[15] D. Kahneman, *Thinking, Fast and Slow*, Penguin, 2012.

[16] S. Amari, *Neuroscience of the Mind: Implications from Biological Psychiatry (Brain Science 6)*, T. Kato, Ed. University of Tokyo Press, 2008.

[17] J. LeDoux, *Synaptic Self: How Our Brains Become Who We Are*, Penguin Books, 2003.

[18] T. Ono and H. Nishijo, "Neural mechanism of intelligence, emotion, and intention," *Brain and Nerve*, vol. 60, no. 9, pp. 995-1007, September 2008.

[19] K. Miyazaki, M. Yamamura, and S. Kobayashi, "MarcoPolo: A reinforcement learning system considering tradeoff exploration and exploitation under Marcovian environments," in *Proc. the 4th International Conference on Fuzzy Logic, Neural Nets and Soft Computing (IIZUKA'96)*, 1996, pp. 561-564.

[20] K. Miyazaki, M. Yamamura, and S. Kobayashi, "MarcoPolo: A reinforcement learning system considering tradeoff exploitation and

exploration under Markovian environments," *Journal of Japanese Society for Artificial Intelligence*, vol. 12, no. 1, pp. 78-89, November 1997.

[21] K. Miyazaki, K. M. Yamamura, and S. Kobayashi, "On the rationality of profit sharing in reinforcement learning," in *Proc. the 3rd International Conference on Fuzzy Logic, Neural Nets and Soft Computing*, 1994, pp. 285-288.

[22] K. Miyazaki, M. Yamamura, and S. Kobayashi, "A theory of profit sharing in reinforcement learning," *Journal of Japanese Society for Artificial Intelligence*, vol. 9, no. 4, pp. 580-587, July 1994.

[23] K. Miyazaki, M. Yamamura, and S. Kobayashi, "k-Certainty exploration method: An action selector to identify the environment in reinforcement learning," *Artificial intelligence*, vol. 91, no. 1, pp. 155-171, March 1997.

[24] R. Brooks, "A robust layered control system for a mobile robot," *Robotics and Automation*, vol. 2, no. 1, pp. 14-23, March 1986.

[25] K. Miyazaki, "Proposal of an exploitation-oriented learning method on multiple rewards and penalties environments and the design guideline," *Journal of Computers*, vol. 8, no. 7, pp. 1683-1690, July 2013.

[26] P. Stone, R. S. Sutton, and G. Kuhlamann, "Reinforcement learning toward RoboCup soccer keepaway," *Adaptive Behavior*, vol. 13, no. 3, pp. 165-188, September 2005.

[27] T. Watanabe, K. Miyazaki, and H. Kobayashi, "A new improved penalty avoiding rational policy making algorithm for keepaway with continuous state spaces," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, no. 6, pp. 675-682, November 2009.

[28] S. Kuroda, S. K. Miyazaki, and S. Kobayashi, "Introduction of fixed mode States into online reinforcement learning with penalty and reward and its application to waist trajectory generation of biped robot," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 16, no. 6, pp. 758-768, September 2012.

[29] K. Miyazaki and M. Ida, "Proposal and evaluation of the active course classification support system with exploitation-oriented learning," *Lecture Notes in Computer Science*, pp. 333-344, 2012.

Kazuteru Miyazaki was born in Kanagawa, Japan, on 17 May, 1967. He received the B.Eng from Meiji University in 1991, and the M.Eng and D.Eng degrees from Tokyo Institute of Technology in 1993 and 1996, respectively. He is currently an associate professor in the National Institution for Academic Degrees and University Evaluation. His research interests include artificial intelligence, machine learning, reinforcement learning, multi-agent systems, text mining, and consciousness systems. He is a member of the Japanese Society for Artificial Intelligence (JSAI), the Society of Instrument and Control Engineers (SICE), Information Processing Society of Japan (IPSJ), the Japan Society of Mechanical Engineers (JSME), the Robot Society of Japan (RSJ), and Japanese Association of Higher Education Research.