2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Multimodal Deep Q-Network for Environmental Adaptation of robotized plants*

Ryo Miwaura[1] and Eri Sato-Shimokawara[1]

Abstract— **Though keeping pets increases communication among family and colleagues, it has challenges, such as allergies and the need for environmental management. As an alternative, we propose the robotized plant, designed to enhance group communication through nurturing activities. We posit that environmental adaptation is essential for the robotized plant to coexist with its caretakers over extended periods. To achieve this, we aim to optimize its vocalization behavior by considering both internal and external states using a Multimodal Deep Q-Network (DQN). This paper evaluates the feasibility of environmental adaptation by analyzing the learning outcomes of the proposed system under various simulated conditions.**

I. Introduction

Keeping pets at home or in the office increases communication within communities, such as with family and colleagues. However, this practice presents challenges like allergies and restrictions on pet ownership due to facility rules, which are notable issues related to environmental management. To overcome such challenges, integrated communication robots could be considered; however, keeping motivated in communication is difficult due to the lack of visible growth in the robots.

To address these problems, we have developed the robotized plant 'Comulop [1]', which integrates a plant with a communication robot, as illustrated in Figure 1. Growing plants is more managable in terms of environmental maintenance than keeping pets, and unlike communication robots, plants provide a visible sense of growth. By having small groups grow the plant, we aim to increase communication within those groups.

In our previous research, we developed modules for detecting people entering a room using facial recognition cameras and for recognizing the behavior of small groups using skeletal data [2]. Based on these modules, our next goal is to improve interactions between robotized plants and growers.

In this research, we focused on Bowlby's internal working models, initially developed within his attachment theory for infants and adapted for application in Human-Robot Interaction (HRI). We hypothesized that enabling the robot to acquire interactions suited to its environment based on the internal working models would improve the quality of interactions and the impression of the robot to humans. Some research aimed at designing HRI with a daily life service robot. By considering the robot's external states, such as the conditions of the surrounding people, these robots are designed to facilitate more interactions. In addition to HPI, the internal states of plants, such as soil moisture levels, must be considered.

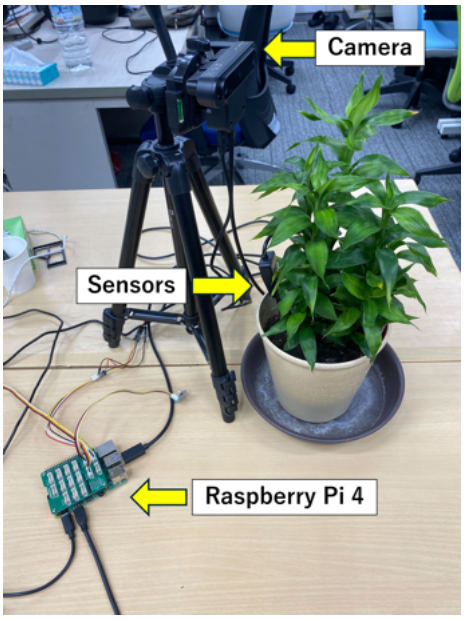


Fig. 1. Prototype of the Robotized Plant

Therefore, this research aims to develop a robotized plant that adapts to its environment by seeking interaction with people and considering its internal and external states. In this paper, we construct a system that uses Multimodal Deep Q-Networks (MDQN) to learn action selection by considering its internal and external states and evaluate this system through simulations.

II. Related Work

A. Integration of Plants and Robots

Numerous studies and products have focused on fusing robots with plants to facilitate HPI [3], [4], [5], [6]. In PotPet [3], built-in sensors detect external and internal conditions such as ambient light levels and soil moisture. The four-wheel drive plant robot then provides feedback to the grower through its movements. In Famileaf [4], interactions with the grower are facilitated by posting the plant's status on social networking.

Our Comulop interacts through vocalization. It captures the attention of nearby growers and encourages communi-

*This work was supported by JSPS KAKENHI Grant Number JP23K11287.

[1]Ryo Miwaura and Eri Sato-Shimokawara are with Graduate School of Systems Design, Computer Science, Tokyo Metropolitan University, 6-6, Asahigaoka, Hino, Tokyo 191-0065, Japan miwaura-ryo@ed.tmu.ac.jp, eri@tmu.ac.jp

cation among them. However, the timing of vocalizations is crucial to eliciting positive reactions. If vocalizations occur without proper timing, growers may perceive them as annoying and disruptive, resulting in a negative impression. Thus, we primarily need to develop a system that vocalizes at appropriate timing to coexist with growers for an extended period of time and increase communication among them.

B. Environment-Adaptive Robots

To develop a robot that adapts to its environment, we focused on Bowlby's attachment theory [7], [8]. In his research, Bowlby introduced the concept of internal working models. According to his model, children internalize information from two perspectives: 'Is the other person responsive to my needs?' and 'Am I worthy of being helped?' Subsequently, using the internalized information, children engage in attachment behaviors tailored to this information to secure protection effectively. Based on this theory, robots that coexist with humans must acquire interactions suited to their specific environments. To achieve this, we selected reinforcement learning that gradually adapts to its environment.

Qureshi et al. [9], [10] optimized robot behavior selection using MDQN with inputs from grayscale and depth images of approaching humans. In this study, we use MDQN with inputs from the internal and external states of the robotized plant to optimize its behavior.

III. The Proposed Method

A. Multimodal DQN

Deep Q-Network (DQN), which can interpret complex environments such as HRI, is commonly used for reinforcement learning. MDQN [9] extends DQN by incorporating multiple modalities, allowing learning from diverse perspectives. MDQN calculates Q-values for input values independently for each modality. Subsequently, these Q-values are integrated to obtain the Fusion Q-value. The Fusion Q-value is calculated by normalizing each Q-value and then taking their average. Finally, it selects an action by choosing the one with the highest value from the Fusion Q-value, and ϵ-greedy policy is employed to incorporate exploratory actions with a certain probability.

In this study, as shown in Figure 2, we derive the Fusion Q-value from the internal state stream obtained through sensing and the external state stream obtained from the camera module.

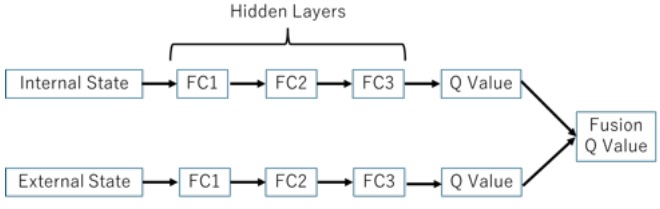


Fig. 2. Architecture of Multimodal DQN

We use a multilayer perceptron as the neural network. It consists of three hidden layers, each with 256 nodes. The activation function employed is the tanh function. The learning rate is set to 0.0001, and the discount factor is 0.9.

1) Internal State: The internal states include 'soil moisture level' and 'the day's foliar watering count', which are measured by sensors and used as input values. Foliar watering is the act of spraying water directly onto the leaves.

The action of the robotized plant is either 'vocalizing' or 'remaining silent.' In 'vocalizing', it produces non-natural language sounds, which likely give growers the impression of animal-like sounds.

$$R = \begin{cases} \text{normalize}(w) & \text{if the robot can get 'water'} \\ 1 - \frac{3n}{10} & \text{if the robot can get 'foliar watering'} \\ -1.0 & \text{if the robot cannot get care} \\ 0.0 & \text{if the robot remains silent} \end{cases} \quad (1)$$

We defined the reward function as shown in Equation 1. If the robotized plant can get 'water' from the growers after vocalizing, the reward is set to the normalized value of the current moisture level w, where $-1 \leq normalize(w) \leq 1$. When the soil moisture level is low, the reward is positive; when the soil moisture level is high, the reward is negative. We defined this reward to account for the risk of root rot due to overwatering and the possibility of plant wilting due to insufficient watering.

If the robotized plant gets 'foliar watering' from the growers after vocalizing, we set the reward to decrease as the day's foliar watering count n increases. The coefficients of the equation were set with three as the reference point, such that the reward becomes negative if it exceeds three times per day. On the other hand, a negative reward is assigned if the robotized plant doesn't get care from the growers after vocalizing.

2) External State: The external states consist of 'the distance to people' and 'the number of people' detected in eight frames of image data obtained from the camera module.

The actions are the same as the internal state.

$$R = \begin{cases} 1.0 & \text{if the robot can get care} \\ -1.0 & \text{if the robot cannot get care} \\ 0.0 & \text{if the robot remains silent} \end{cases} \quad (2)$$

We defined the reward function as shown in Equation 2. If the robotized plant receives care from the growers after vocalizing, a positive reward is assigned. Conversely, a negative reward is given, if it does not receive care. And if it remains silent, it receives a reward of 0.

B. Learning Procedure

A large amount of experiential data is required when training with DQN. However, in HRI, providing sufficient experimental data for training is challenging. Therefore, we employ the method proposed by AH Qureshi et al. [10] Their proposed method improves learning efficiency by dividing the process into two phases: the Data Generation Phase and the Training Phase.

1) Data Generation Phase: Data generation phase involves conducting HRI to gather data as illustrated in Figure 3.

First, the system receives inputs from the environment and calculates the Q-values using the policy obtained from the current learning process. Next, the action is determined and executed based on the Q-values and the ϵ-greedy policy. Then, the reward is determined based on the human response to the executed action. Finally, the state, action, reward, and subsequent state from this experience are stored in the replay memory.

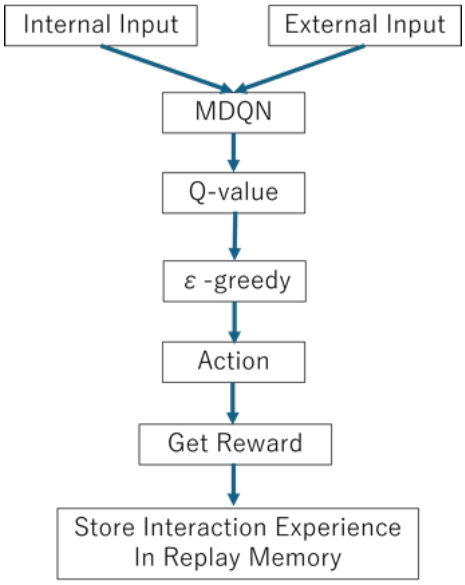


Fig. 3. Data Generation Phase

2) Training Phase: In the training Phase, as illustrated in Figure 4, learning is repeatedly conducted based on the stored experiences. First, experiences stored in the replay memory are sampled randomly. The target Q-value is computed from the sampled experiences. Then, the loss function is computed as the difference between the target Q-value and the current Q-value. Learning is achieved by minimizing this loss function using gradient descent methods.

By dividing the data generation phase and training phase, the additional load on the robot during interactions can be avoided, which could otherwise lead to issues such as freezing or lag.

The data generation phase is performed when people are present, while the training phase is executed when people aren't present.

In our study, the transition between the data generation and the training phases is controlled based on the illumination level of the room, as illustrated in Figure 5. When the room is bright and frequented by people, the data generation phase is conducted to accumulate episodes. When the room becomes dark and empty, the training phase begins for learning. Then, when the room becomes bright, the policy learned then is used as the policy for the data generation phase.

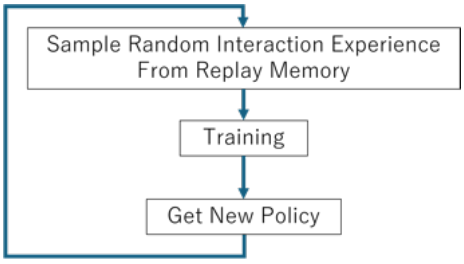


Fig. 4. Training Phase

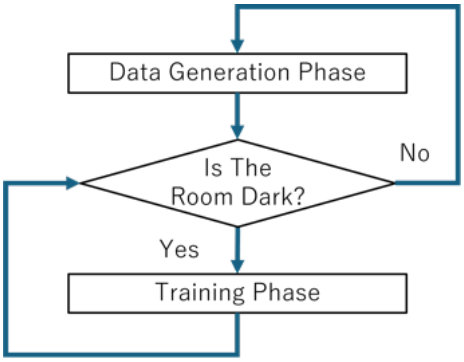


Fig. 5. Interaction Learning Procedure

C. Our Method

In our approach with Multimodal DQN, when normalizing and averaging the internal and the external states, we hypothesized that introducing a bias toward either of these states expresses the robot's personality.

$$Q_{\text{fusion}} = \frac{\alpha Q_{\text{internal}} + \beta Q_{\text{external}}}{\alpha + \beta} \tag{3}$$

Note: α and β are relatively prime natural numbers.

In Equation 3, if $\alpha > \beta$, the internal state is considered heavily. And the robotized plant is expected to become more autonomous. On the other hand, if $\alpha < \beta$, the robotized plant is expected to be more adaptable to its environment and take more cautious actions.

We conducted a simulation to verify whether our method contributes to environmental adaptation.

IV. Simulation

In simulation, we compared the variations in the probability of receiving care from people after vocalization. And the parameters of Equation 3 were set to $\alpha = 2$ and $\beta = 1$.

A. Simulation Setup

The input values were generated randomly for both internal and external states.

1) Internal State Input: The input values for the internal state are the soil moisture level w and the day's foliar watering count $n_{\text{foliar watering}}$, as shown in Equation 4.

The soil moisture level w is generated uniformly within the range of $310 \leq w \leq 780$ based on actual measured values.

Here, due to the specifications of the sensors employed, a smaller value of w corresponds to higher soil moisture content, whereas a larger value of w indicates drier soil conditions.

The day's foliar watering count $n_{\text{foliar watering}}$ is generated as an integer value following a uniform random distribution within the range $0 \leq n_{\text{foliar watering}} \leq 5$.

$$state_{\text{internal}} = \begin{bmatrix} w \\ n_{\text{foliar watering}} \end{bmatrix} \tag{4}$$

2) External State Input: The external state inputs comprised a sequence of distances between the robotized plant and the closest person, obtained from eight frames of image data and the maximum number of people detected across those frames, as shown in Equation 5.

In this simulation, instead of using image data, the distance sequence and number of people are generated using random numbers.

$$state_{\text{external}} = \begin{bmatrix} |x| \\ |x+d| \\ |x+2d| \\ \vdots \\ |x+7d| \\ n_{\text{people}} \end{bmatrix} \tag{5}$$

The initial element x of the sequence of distances is generated as a uniform random variable within the range of 0 to 3 meters. This range of x represents th distance from the desk in the shared space of our laboratory to th entrance of the laboratory. The robotized plant is placed on the desk; this simulation focuses on optimizing the robotized plant's vocalization in response to the entry and exit of growers. And the movement distance d per a frame is generated using a uniform random distribution within the range of $-0.3 \leq x \leq 0.3$ meters. A negative value of d represents the grower's approach, while a positive value signifies that the distance between the robotized plant and the grower is increasing.

The number of people n_{people} is generated as a random number following a uniform distribution in the range of $1 \leq n_{\text{people}} \leq 4$.

3) Determination of Care Probabilities: The probability of receiving care after vocalization p is determined from the external state input values, as described in Equations 6, 7 and 8.

The probability p_1 was set to increase with n_{people}, as shown in Equation 6. The probability p_2 increases, as shown in Equation 7, as the average distance, AvgDistance, derived from the sequence of distance data, decreases.

$$p_1 = \gamma \cdot n_{\text{people}} \tag{6}$$

$$p_2 = 1 - \frac{\text{AvgDistance}}{\delta} \tag{7}$$

$$p = p_1 + p_2 \tag{8}$$

The probability p is determined by summing the probabilities p_1 and p_2.

By varying the parameters γ and δ in Equation 8, we adjusted the probability of receiving care after vocalization.

The probability p is used to determine, as shown in Equation 9, which type of care will be received, or care will not be received.

$$care = \begin{cases} \text{cannot receive care} & \text{with probability } 1-p \\ \text{“water”} & \text{with probability } \frac{p}{2} \\ \text{“foliar watering”} & \text{with probability } \frac{p}{2} \end{cases} \tag{9}$$

B. Condition

The probability of receiving care after vocalization is adjusted by considering the averages of the input values and modifying the parameters in Equation 6 and 7 accordingly. And the system was trained for 100,000 iterations.

Condition 1

In Condition 1, the robotized plant was set to receive care with an average probability of 50% after vocalizing. The parameters of Equation 6 and 7 were set to $\gamma = 0.1$ and $\delta = 2$.

Condition 2

In Condition 2, the robotized plant was set to receive care with an average probability of 75% after vocalizing. The parameters of Equation 6 and 7 were set to $\gamma = 0.15$ and $\delta = 2.4$.

C. Result and Discussion

Figure 6 illustrate the progression of average rewards under Conditions 1 and 2, respectively. Figure 7 depict the progression of loss rates under Conditions 1 and 2, respectively. And Table I shows the number of vocalizations along with the average input values when 100 pre-prepared input datasets were provided.

Looking at Figure 6, the average rewards are increasing under both Condition 1 and Condition 2. It indicates that the agent is learning to take better actions and increasingly adapting to its environment.

On the other hand, as shown in Figure 7, the external state DQN showed a decrease in loss rate, suggesting that the learning of the MDQN was progressing. However, there was little change in the loss rate for the internal state. Since the loss rate of the internal state DQN remains unchanged, it suggests that the agent's learning is still unstable or that exploratory behavior is ongoing. In this case, it is likely that the agent has not yet reached an optimal policy and is still in the learning phase.

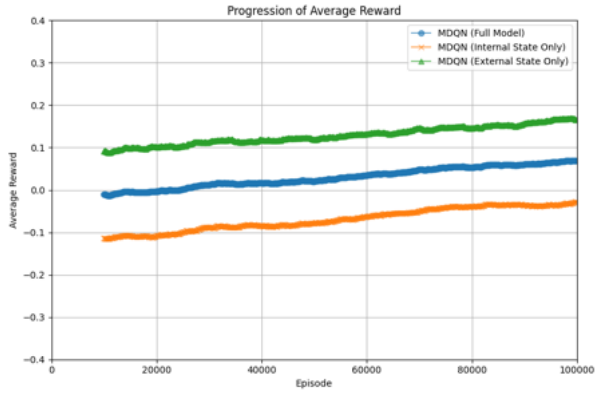
Additionally, a comparison of Figures 6(a) and 6(b) reveals that the final average reward in Condition 2 is higher than in Condition 1. Furthermore, comparing Table I, it is evident that the frequency of vocalizations in Condition 2 is nearly double that in Condition 1. When considering the average input values during vocalization, it becomes clear that vocalization is more likely to occur in Condition 2.

These observations suggest that the robotized plant increases its vocalizations when it is in an environment where

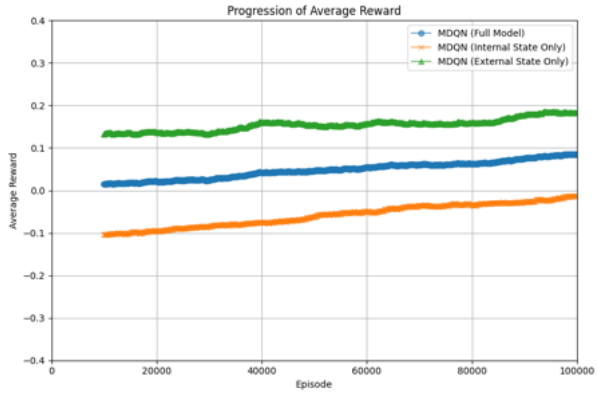
TABLE I

NUMBER OF VOCALIZATIONS AND THE MEAN INPUT VALUES DURING VOCALIZATION FOR 100 DATA POINTS

Probability	Vocalization Count	Average Distance	Average Number of People	Average Moisture	Average Number of Foliar Waterings
50%	27	0.8652	2.5556	637.4250	2.2222
75%	47	1.2138	2.6596	622.2690	2.3830

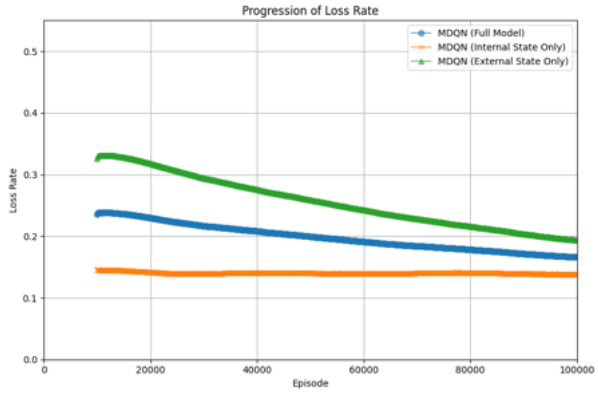


(a) Condition 1

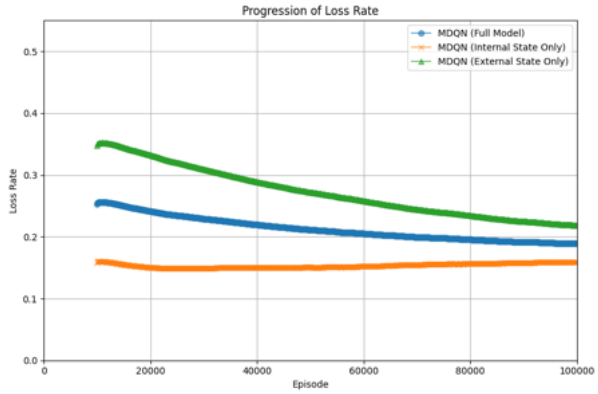


(b) Condition 2

Fig. 6. Progression of Average Reward



(a) Condition 1



(b) Condition 2

Fig. 7. Progression of Loss Rate

care is more likely to be received, and limits vocalization to only certain instances when care is less likely.

Therefore, it can be concluded that the adaptation of the robotized plant to its environment using MDQN is feasible in simulation.

V. CONCLUSIONS

In this study, we implemented the proposed system to enable the robotized plant to adapt to its environment within a simulated setting. These findings suggest that effective symbiosis between robots and their environment is feasible, potentially leading to improved adaptability of the robot.

However, there remains a significant challenge concerning instability in learning the internal state. Specifically, insufficient convergence of loss rates under certain environmental conditions indicates the need for further examination to enhance the stability of the learning algorithm.

Future research will involve deploying the robotized plant equipped with the proposed system in multiple real-world environments to assess its practical viability by observing behavioral changes. Additionally, it is essential to conduct a detailed investigation into how variations in the parameters included in Equation 3 affect the ”character expression” of the robotized plant. This research could enable differentiation among robots by adjusting parameters, making it possible for caregivers to perceive a sense of individuality or lifelike qualities when multiple robotized plants are placed in the same environment. Such advancements are expected to contribute to more effective symbiosis.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number JP23K11287.

REFERENCES

[1] Ryo Miwaura, Koki Baba, and Eri Sato-Shimokawara. Botanical communication robot based on emotional model considering external factors. *Proceedings of the Fuzzy System Symposium*, 39:536–541, 2023 (in Japanese).

[2] Koki Baba, Ryo Miwaura, and Eri Sato-Shimokawara. Small group behavior recognition with skeleton data for human-robot interaction. In *SICE System Integration Division SI2023*, pages 3124–3127. SICE, 2023 (in Japanese).

[3] Ayumi Kawakami, Koji Tsukada, Keisuke Kambara, and Itiro Siio. Potpet: pet-like flowerpot robot. In *Proceedings of the fifth international conference on Tangible, embedded, and embodied interaction*, pages 263–264, 2010.

[4] Nagisa Ishizumi and Manabu Gouko. Famileaf: Flowerpot robot for dementia prevention. In *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pages 1–2. IEEE, 2019.

[5] Van Oers, E. Living light - lamp. `https://livinglight.info/about/`, 2016.

[6] Steiner Helene, Johns Paul, Roseway Asta, Quirk Chris, Gupta Sidhant, and Lester Jonathan. Project florence. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, page 1415, 05 2017.

[7] Inge Bretherton. The origins of attachment theory: John bowlby and mary ainsworth. In *Attachment theory*, pages 45–84. Routledge, 2013.

[8] Nicholas Rabb, Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. An attachment framework for human-robot interaction. *International journal of social robotics*, pages 1–21, 2022.

[9] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Robot gains social intelligence through multimodal deep reinforcement learning. In *2016 IEEE-RAS 16th international conference on humanoid robots (humanoids)*, pages 745–751. IEEE, 2016.

[10] Ahmed Hussain Qureshi, Yutaka Nakamura, Yuichiro Yoshikawa, and Hiroshi Ishiguro. Intrinsically motivated reinforcement learning for human–robot interaction in the real-world. *Neural Networks*, 107:23–33, 2018.