Received: 22 April 2024 | Accepted: 5 February 2025

DOI: 10.1002/jeab.70004

RESEARCH ARTICLE

Journal of the Experimental Analysis of Behavior

SEAB

Of rats and robots: A mutual learning paradigm

Oguzcan Nas | **Defne Albayrak** | **Gunes Unal**

Behavioral Neuroscience Laboratory, Department of Psychology, Boğaziçi University, Istanbul, Turkey

Correspondence
Gunes Unal, Behavioral Neuroscience Laboratory, Department of Psychology, Boğaziçi University, 34342 Istanbul, Turkey.
Email: gunes.unal@bogazici.edu.tr

Funding information
Boğaziçi Üniversitesi, Grant/Award Number: 22B07M2

Editor-in-Chief: Suzanne H. Mitchell
Handling Editor: Cynthia Pietras

Abstract

Robots are increasingly used alongside Skinner boxes to train animals in operant conditioning tasks. Similarly, animals are being employed in artificial intelligence research to train various algorithms. However, both types of experiments rely on unidirectional learning, where one partner—the animal or the robot—acts as the teacher and the other as the student. Here, we present a novel animal–robot interaction paradigm that enables bidirectional, or mutual, learning between a Wistar rat and a robot. The two agents interacted with each other to achieve specific goals, dynamically adjusting their actions based on the positive (rewarding) or negative (punishing) signals provided by their partner. The paradigm was tested in silico with two artificial reinforcement learning agents and in vivo with different rat–robot pairs. In the virtual trials, both agents were able to adapt their behavior toward reward maximization, achieving mutual learning. The in vivo experiments revealed that rats rapidly acquired the behaviors necessary to receive the reward and exhibited passive avoidance learning for negative signals when the robot displayed a steep learning curve. The developed paradigm can be used in various animal–machine interactions to test the efficacy of different learning rules and reinforcement schedules.

KEYWORDS
animal–robot interaction, mutual learning, operant conditioning, optimization, reinforcement

Introduction

The use of machinery in the behavioral school of psychology dates back to Edward Thorndike's seminal puzzle box experiments, where confined cats had to learn a simple task, such as pressing a lever, to escape from the puzzle box (Thorndike, 1898). This was followed by B. F. Skinner's work on operant conditioning in which he added the concept of reinforcement (Skinner, 1935) to Thorndike's law of effect (Thorndike, 1927) and developed the operant conditioning chamber known as the Skinner box (Skinner, 1932). Although these early experiments used handmade machinery to study behavior, recent advances in technology have resulted in the integration of robotics in behavioral research (Li et al., 2022). Unlike operant chambers, which measure behavior in static environments, robots enable the creation of dynamic, interactive settings, allowing for the study of more complex behaviors.

In recent years, robot use has gained popularity in various behavioral experiments that investigate visual perception (Frohnwieser et al., 2016), courting behavior (Klein et al., 2012), social learning (Frohnwieser et al., 2016), cooperation (Krause et al., 2011), and self-organization (Romano et al., 2018). Although robotics is increasingly used to study animal behavior (Li et al., 2022), different animals including rats, sheep, frogs, ducks, and zebrafish (Romano et al., 2018) are used in artificial intelligence research to train and provide feedback to different algorithms (Hassabis et al., 2017). In both types of experiments, the interactions between animals and robots are unidirectional: either robots replace the Skinner box to elucidate more complex animal behaviors, such as helping and predation (Quinn et al., 2018; Rundus et al., 2007), or animals constitute passive agents that facilitate kinesthetic learning by the robot (Peng et al., 2020). In this proof-of-concept study, we present a novel rat–robot interaction

paradigm that enables mutual learning between the partners. A Wistar rat and a robot interact with each other to achieve specific goals by simultaneously adapting their own actions in response to the actions of their partner. We tested this paradigm in silico with two artificial reinforcement learning agents and in vivo by using different rat–robot pairs.

Robots facilitate the creation of dynamic, interactive environments that better simulate natural conditions, offering considerable advantages over static setups. They enable interaction paradigms and the study of adaptive behaviors (Miklósi & Gerencsér, 2012), allowing researchers to investigate complex phenomena such as social interactions and species-specific behaviors that may be difficult to explore in conventional Skinner boxes. For example, robots can train a rat to navigate a complex maze, a task the rat might not accomplish on its own (Gianelli et al., 2018). Robots can also be designed to mimic conspecifics or predators, eliciting species-specific behaviors (Romano et al., 2018). Biomimetic robots, which morphologically resemble the target organism, have proven especially useful for this purpose. These robots were used in ichthyology to study interactions between individual fish when placed within a shoal of guppies (Bierbach et al., 2018). In another study, a biomimetic robot not only integrated into a biological group of zebrafish (Cazenille et al., 2018) but also modulated and directed them by displaying species-specific leadership behavior (Chemtob et al., 2020). Biologically inspired robots have also been used to study predator threat and avoidance behavior in squirrels (Rundus et al., 2007), social interaction in cockroaches (Asadpour et al., 2006), and operant learning in rodents (Xie et al., 2023). Highly capable and sophisticated biomimetic robots have been developed in recent years to study behavioral modulation and social interaction parameters among laboratory animals (Shi, Ishii, Sugahara, et al., 2015).

The use of robots in social interaction research can take place in both experimental settings and the natural environment of the animal. Robots can be employed to perform intricate actions that elicit specific social behaviors in rodents. In one study, a rat-like robot modulated the social interaction dynamics between multiple rats, leading to increased agonistic behavior and locomotion (Shi, Ishii, Tanaka, et al., 2015). In another study, individual rats displayed social interaction with robots by receiving and returning favors (Quinn et al., 2018). The rats were then tasked with rescuing the robots, and they prioritized the "helpful" robots—those that had previously helped them—over the unhelpful ones (Quinn et al., 2018).

Robots can be enhanced with artificial intelligence to perform their dedicated actions in experimental settings (Shi, Ishii, Kinoshita, Konno, et al., 2013). For instance, an e-puck mobile robot, programmed with an action-selection model operating through sensory signals, was used to trigger social behavior in rats (Del Angel Ortiz et al., 2016). Similarly, the rat-like robot used in the social interaction study by Shi, Ishii, Sugahara, et al. (2015) was capable of transmitting stress signals or displaying friendly behavior to condition the animals (Shi, Ishii, Kinoshita, Takanishi, et al., 2013).

Behavioral paradigms and cognitive models are translated into robotics to enhance the capabilities of robots, enabling them to perform more sophisticated functions. A pattern generation algorithm that incorporates shaping-like behavior was used to autonomously condition a rat to press a lever on the robot (Ishii et al., 2006). These robots can be remotely controlled or programmed to perform preset actions, interacting with animals through a series of events (Abdai et al., 2018). Semiautonomous robots can also teach rats complex multistep tasks through shaping (Ishii et al., 2006). In an earlier study by the same group, remotely controlled robots could train rats to press a moving lever (i.e., located on the robot) to obtain food, showing successful operant conditioning (Ishii et al., 2003).

The interaction between animal testing and robotics is not limited to the use of robots as experimental tools. Artificially intelligent robots possess the ability to observe animals and learn from their behavior, enabling them to adjust and modify their own actions accordingly (Peng et al., 2020). Furthermore, these artificially intelligent robots can observe the consequences of their actions on animals and adjust their future behavior to maximize rewards (Son et al., 2014). These robots commonly employ a fundamental machine-learning paradigm called reinforcement learning to execute their functions (Abbeel & Ng, 2004; Buşoniu et al., 2010; Kober et al., 2013). Within the reinforcement learning algorithms, Q-learning emerges as a practical tool for piloting experiments that study behavior in dynamic environments. This feature makes it a valuable algorithm in studies where an agent must interact with either the environment or another agent (Sutton & Barto, 2018). It must be noted that there is an analogy between reinforcement learning of artificially intelligent agents and the concept of reinforcement in operant conditioning (Skinner, 1935), as both phenomena refer to strengthening of a particular behavior in response to observing the positive effects of that behavior on the environment (Ludvig et al., 2010). Artificially intelligent agents using reinforcement learning can learn to exhibit species-specific behaviors like sniffing or grooming (Xie et al., 2023), imitate animal motion as demonstrated in a simulation environment (Xie et al., 2022), and adapt to different reinforcement intervals (Silveira et al., 2023). Nevertheless, existing research (Gianelli et al., 2018; Ishii et al., 2006; Peng et al., 2020; Shi, Ishii, Kinoshita, Takanishi, et al., 2013) has concentrated on assessing and maximizing the learning performance of either the experimental animal or the robot, treating the other agent merely as an experimental tool.

Here, we present a new framework in which a rat and a robot simultaneously learn from each other by

observing the effects of their own actions on the other party. We first tested the paradigm in a virtual setting under different conditions. The in silico experiments serve as a detailed tutorial on how reinforcement learning and the Q-learning algorithm can be used to study complex multiagent behaviors. Subsequently, we conducted in vivo experiments, where different pairs of rats and robots interacted for reward maximization. In successful trials, the rat learned the rules of the paradigm by observing the robot's actions and the rat's behavior provided feedback to the robot, altering its signaling process and leading to mutual learning.

IN SILICO EXPERIMENTS

Methods

Apparatus

The simulation environment was developed in Python to test virtual interactions between two reinforcement learning agents in different settings and scenarios. We developed a simple simulation engine to track and manipulate the agents' positions and states based on their actions, applying rules for reward and punishment. The simulation code is provided in Supporting information S1. The virtual environment was a 10×10 square grid, comprising 100 positions identified by their coordinates (x, y), ranging from 0 to 9. A stationary food dispenser (feeder) was placed in the top-right corner of the grid to deliver rewards. Agents moved between tiles in discrete steps and were restricted to the boundaries of the virtual grid. Each state within the grid was represented by a combination of the responder's current position, the signaler's signal, and whether the responder was near the signaler or the food dispenser. The responder (virtual rat) occupied a single tile, whereas the signaler (virtual robot) was positioned at the center of a two-by-two square area, designated as the *signaler zone*. The interaction zone, also a two-by-two square area, moved with the signaler and defined the space where interactions between the agents occurred. Each instance of the responder entering the interaction zone was recorded as an interaction between the two agents.

Procedure

In the simulation, timing and actions were measured in discrete units and steps, with each action by the signaler or responder considered a single turn. Each trial started when the signaler, located at the center of the grid, made a decision to present either a positive signal or a negative signal and then moved two tiles to the right or left. The signaler moved to the left following the negative signal and to the right after the positive signal. Once the signaler's movement was completed, the responder started moving from its initial position in the lower-left corner of the virtual grid. The positive signal indicated that a reward (+1,000 points) would be given to the responder if it performed the desired action associated with the positive signal. Conversely, the negative signal indicated that the responder would receive a punishment (−1,000 points) if it misinterpreted the negative signal as positive and performed the corresponding action. These point values were set to ensure they were sufficiently distinct to facilitate rapid learning and discourage point accumulation through random wandering. The signaler also received a reinforcement of +1,000 points for successful interactions but was not penalized for incorrect outcomes (i.e., 0 points). Each trial concluded when an interaction occurred or after 200 turns.

The desired action for the responder in the presence of a positive signal was to first enter the interaction zone (i.e., perform an interaction) and then move to the feeder tile within 30 moves. Reinforcement of the responder was contingent on this response following a positive signal, akin to the discriminative stimulus for reinforcement (S^{Dr}). Moving next to the feeder in the absence of a positive signal did not lead to a reinforcement or a punishment. The responder's behavior was punished when it entered the interaction zone following a negative signal. When the signaler presented a negative signal, the responder had no chance of receiving a reward and the best outcome for it would be to avoid punishment by not approaching the signaler. The negative signal therefore constituted a discriminative stimulus for punishment (S^{Dp}).

The signaler, in turn, was rewarded each time the responder entered the interaction zone regardless of the number of moves or the type of preceding signal. Thus, interaction with the responder constituted the reinforcement of the signaler's behavior. There was no punishment for the signaler, and attracting the responder was the only factor driving the signaler to modify its behavior. However, the signaler did not know the nature of its signals at the beginning of the experiment and had to determine which signal to present by observing the effects of its signals on the responder (i.e., whether a signal led to an interaction and subsequent reinforcement). Both the responder and the signaler were reinforcement learning agents. Their probability of performing a particular action in a given situation increased when that action resulted in a reward or led to a new situation that was closer to a reward. This learning process was achieved by the Q-learning algorithm, as explained below.

Agents and learning rules

Both agents used the tabular Q-learning algorithm (Clifton & Laber, 2020) for learning and the epsilon-greedy algorithm (Zawadzki et al., 2014) for exploration.

The epsilon-greedy algorithm was selected due to its efficient balance of exploration and exploitation in contrast to other adaptive algorithms such as softmax or upper confidence bound, which require parameter tuning that might not accurately reflect stochastic decision making (Sutton & Barto, 2018). The agents had Q tables that contain the Q value for all the possible actions in all discrete states. The responder could display six actions. These include (1) moving up, (2) moving down, (3) moving to the left, (4) moving to the right, (5) waiting, and (6) accessing the feeder. The responder could occupy any of the 100 tiles within the virtual grid. It received input from the signaler (i.e., the positive or negative signal) and tracked whether it had been within the interaction zone during the last 30 moves. This gives 400 distinct states—that is, 2 (signal states: positive or negative) × 2 (interaction zone states: within the last 30 turns or not) × 100 (tile positions) = 400 states). With six possible actions available in each state, the responder maintained a total of 2,400 Q values. In contrast, the signaler had only one state with two actions: giving a positive or a negative signal. As a result, it learned two Q values to choose between these states.

Before each action, a random number between zero and one was drawn from a uniform distribution and compared with the epsilon value. If it was less than epsilon, the agent took a random action. When a random action was taken, each action had an equal probability of being performed. Thus, the chance (percentage) of behaving randomly was set by the epsilon value, which decreased with each trial with the following formula:

$$\epsilon_t = \epsilon_{initial} \cdot \lambda_\epsilon{}^i, \quad (1)$$

where ϵ_t is the current epsilon value, $\epsilon_{initial}$ is the epsilon value at the start of the experiment, λ_ϵ is the epsilon discount factor, and i is the current trial number.

When the agents did not take a random action, they carried out the action with the highest Q value for that given state. Following each action, the agent updated the Q value of the last action in the previous state according to the Q values of their new state or the reward they received in the new state by using this formula (Watkins, 1989; Watkins & Dayan, 1992):

$$Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma \max_{a'} Q(s',a') - Q(s,a)\right], \quad (2)$$

where $Q(s,a)$ is the Q value of taking the action a in state s, α is the learning rate, r is the reward from the action taken, γ is the discount factor, and $\max_{a'} Q(s',a')$ is the Q value of the action in the new state with the highest Q value. The learning rate was initially set to 0.15 for the responder and 0.80 for the signaler. As the responder had multiple tasks, its learning rate was chosen to promote a more gradual and stable learning process, minimizing the influence of short-term changes in reward and punishment. In contrast, the signaler's learning rate was set higher, as it had only two actions, allowing it to quickly adjust to the responder's behavior. The discount factor was set to 0.95 for both agents to prioritize long-term consequences and ensure consistency.

Simulation scenarios

In silico experiments were conducted with four responder–signaler pairs assessing four different scenarios. Each pair was simulated for 10,000 trials, and each simulation was repeated eight times. The first pair (Responder–Signaler Pair 1) was the control condition, for which no external manipulation was made. The learning performance of the responder in the second pair (Responder–Signaler Pair 2) was disrupted by reducing its learning rate to 5% of that in Pair 1. Responder–Signaler Pair 3 tested the scenario in which the signaler behaved completely randomly while the responder behaved as in the control condition. The final scenario, Responder–Signaler Pair 4, tested the reverse condition, where the responder behaved completely randomly and the signaler had the same parameters as in the control condition. To achieve random behavior in Pair 3 (random actions by the signaler) and Pair 4 (random actions by the responder), the epsilon value and the epsilon discount factor were set to 1. The net amount of reinforcement for each trial was calculated for the responders by subtracting the total punishment they received from their total reward.

Results

Data handling

Learning performance of the responders was calculated by revealing their net reinforcement (i.e., punishment points subtracted from reward points) across trials. A moving average of 500 values was calculated for each simulation of 10,000 trials to smooth out the early random fluctuations of the reinforcement points in simulations. This resulted in 9,501 moving averages for eight simulations of four groups each, calculated separately for the responder and signaler. Figure 1 shows the moving average values of net reinforcement points for the responder and signaler across the four pairings. To prevent p value deflation caused by the large number of values, we binned the data into five bins for each simulation. The five bins included the following moving average scores: (0, 2,000] (Bin 1); (2,000, 4,000] (Bin 2); (4,000, 6,000] (Bin 3); (6,000, 8,000] (Bin 4); (8,000, 9,501] (Bin 5). The uneven size of the last bin was disregarded because it would minimally affect the

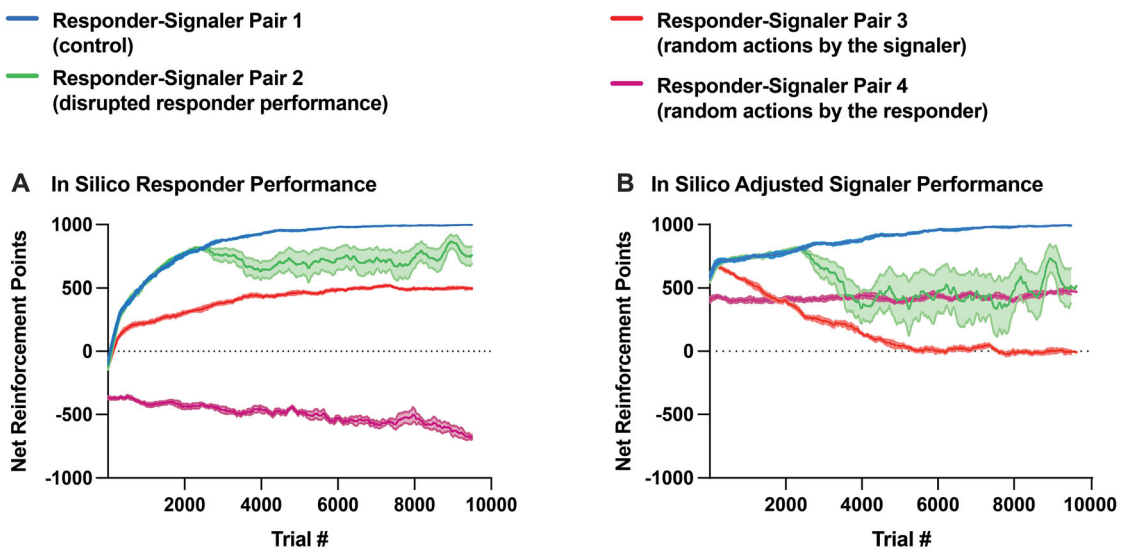

FIGURE 1 Moving averages of net reinforcement points of the virtual rat (Responder) and the virtual robot (Signaler) across 10,000 trials in different in silico Responder–Signaler pairs (color-coded). Both panels illustrate averaged values derived from the eight simulations conducted for each pair. Panel A: Each responder received 1,000 points for a reward and − 1,000 points for a punishment. Net reinforcement was calculated by subtracting punishment points from reward points. Panel B: The signaler received 1,000 points as a reward, with its reinforcement values adjusted to align with the scale used in Panel A. The light-colored areas around the lines represent the 95% confidence interval.

analysis due to the high number of values in each bin. Data handling and analyses were performed in RStudio Version 2023.06.1 + 524.

Clustered regression model

To account for the different simulations and based on the assumption that successive data within one simulation are more similar than data between different simulations (Graubard & Korn, 1994), we clustered the data by simulation for the analysis to obtain clustered standard errors. We then conducted a clustered regression analysis for each pairing of both agents, using heteroscedasticity-robust covariance estimates to account for potential relationships within the simulations. The following regression model was fit using the mean moving average values, which represented the entire bin as the dependent (response) variable and bin as the independent (predictor) variable:

$$Y = \beta_{B_1} + I_{B_2}\beta_{B_2} + I_{B_3}\beta_{B_3} + I_{B_4}\beta_{B_4} + I_{B_5}\beta_{B_5}, \quad (3)$$

using the following formula for assigning the moving average values to their respective bins for the regression equation:

$$I_{B_i} = \begin{cases} 1, & X \in B_i \\ 0, & X \notin B_i \end{cases} \quad (4)$$

while

$$B_1 \ni (0, 2.000],$$

$$B_2 \ni (2.000, 4.000],$$

$$B_3 \ni (4.000, 6.000],$$

$$B_4 \ni (6.000, 8.000],$$

$$B_5 \ni (8.000, 9.501],$$

where Y is the mean moving average value, B_i is the respective bins, and I_{B_i} is the indicator variable for the assignment of the trials to their respective bins.

We then calculated the clustered standard errors of the coefficients through the robust variance–covariance matrix type one (Hayes & Cai, 2007) with the following formula:

$$SE_i = \sqrt{V_{ii}}, \quad (5)$$

where SE_i refers to the clustered standard errors of the coefficients and V represents the element of the heteroscedasticity-consistent covariance matrix. Appendix A (Table A1) and Appendix B (Table B1) show the results of the regression analysis including the coefficients of the bins, clustered standard errors, t values, respective p values, and adjusted R^2 values for the responder and signaler, respectively. The value for the coefficients of each model summed with the respective intercept represents the corresponding bin mean, allowing for a direct interpretation. All coefficients associated with the bins in

four pairings were statistically significant ($p < .001$, one-sample t test). For the responder, the adjusted R^2 values for the Responder–Signaler Pair 1 (control) and Responder–Signaler Pair 3 (random actions by the signaler) were .779 and .800, respectively, indicating that the bin divisions effectively capture the dynamics of the simulated behavior and provide a good fit for both pairings. For Responder–Signaler Pair 2 (disrupted responder performance) and Responder–Signaler Pair 4 (random actions by the responder), the adjusted R^2 values were .171 and .473, respectively, reflecting the randomness and the unstable patterns in learning performance (Figure 1). For the signaler, the adjusted R^2 values for Responder–Signaler Pair 1 and 3 were .893 and .903, respectively, indicating a good fit and a stable pattern in the data. However, the adjusted R^2 values for Responder–Signaler Pair 2 and 4, were .102 and .107, respectively, similar to the responder performance in these pairs.

Statistical testing

We conducted individual independent-samples t tests for the regression coefficients to compare the different bins or different points of the simulation within and between different responder–signaler pairs. As mentioned above, we compared the means of different bins, as the coefficients added to the intercept reflected these means. We used the Benjamini–Hochberg false discovery rate correction for all multiple comparisons. To assess the performance of the virtual agents in the pairs, we conducted four individual t tests and compared the different bins with each other in the order below:

1. First 2,000 trials (Bin 1)–trials between 2,001 and 4,000 (Bin 2),
2. Trials between 2,001 and 4,000 (Bin 2)–trials between 4,001 and 6,000 (Bin 3),
3. Trials between 4,001 and 6,000 (Bin 3)–trials between 6,001 and 8,000 (Bin 4),
4. Trials between 6,001 and 8,000 (Bin 4)–trials between 8,001 and 9,501 (Bin 5).

Comparison of different trials within responder–signaler pairs

In each pair, we compared the performance of each agent between subsequent trial bins using independent-samples t tests. Statistical results are provided in Appendix C (Table C1, responder) and Appendix D (Table D1, signaler), and the comparison between moving averages of net reinforcement points is shown in Figure 2.

In Responder–Signaler Pair 1, the responder's reinforcement points increased steadily until Bin 5, indicating successful learning. No significant difference was found between Bins 4 and 5, suggesting a ceiling effect (Figure 2A). The signaler's reinforcement points continued to increase in Bin 5, indicating ongoing learning through to the end (Figure 2A). In Responder–Signaler Pair 2, the responder showed initial learning, with reinforcement points increasing between Bins 1 and 2 but then decreasing from Bin 2 to Bin 3, followed by a rebound in the last two bins (Appendix C, Table C1). The signaler's performance worsened after the initial learning phase, with decreased reinforcement points between Bins 1 and 2 and again between Bins 2 and 3 (Appendix D, Table D1). Both agents exhibited stagnation between Bins 3 and 4, followed by an increase in reinforcement points in the final trials (Figure 2B). In Responder–Signaler Pair 3, the responder started with low reinforcement points in the first bin but consistently increased them until Bin 4 (Figure 2C). The signaler initially performed well, but its reinforcement points steadily decreased, reaching near zero by the last trial bin (Figure 2C). In Responder–Signaler Pair 4, the responder, which acted randomly, accumulated negative points in all bins, showing more punishment than reinforcement (Appendix C, Table C1). Punishment points increased across bins, but at a decreasing rate (Figure 2D). The signaler's performance was more stable, with reinforcement points increasing in Bins 2, 4, and 5 relative to those obtained in the previous bin (Figure 2D).

Comparison of different trials between responder–signaler pairs

We compared the trial-specific performance of the virtual agents in each pair with independent-samples t tests. Appendix E (Table E1) and Appendix F (Table F1) show the performance difference between different in silico pairs in different sets of trials.

In the first 2,000 trials, the responder in Pair 3 (testing random signaler actions) earned fewer reinforcement points than both the control and Responder–Signaler Pair 2 (Figure 1A). The responder in Pair 4, which acted randomly, was the only agent to earn negative points, performing significantly worse than both the control and the other pairs (Appendix E, Table E1). The signaler in Pair 1 (control) outperformed those in Pairs 3 and 4, with the signaler in Pair 2 also performing better than the one in Pair 3, reflecting the influence of random signaler behavior (Appendix F, Table F1).

In Bin 2, the responder in Pairs 2 and 3 earned fewer points than the responder in Pair 1, with Pair 2 outperforming Pair 3 but still earning more than the responder in Pair 4, which continued to accumulate negative points (Appendix E, Table E1). As in Bin 1, the signaler in Pair 1 outperformed those in Pair 2. The signaler in Pair 3 performed worse than in Pair 1, and the signaler in Pair 2 earned more points than in Pair 3. The signaler in Pair 4 performed poorly, with random behavior in Pair 3 yielding better results than the signaler in Pair 4 (Appendix F, Table F1).

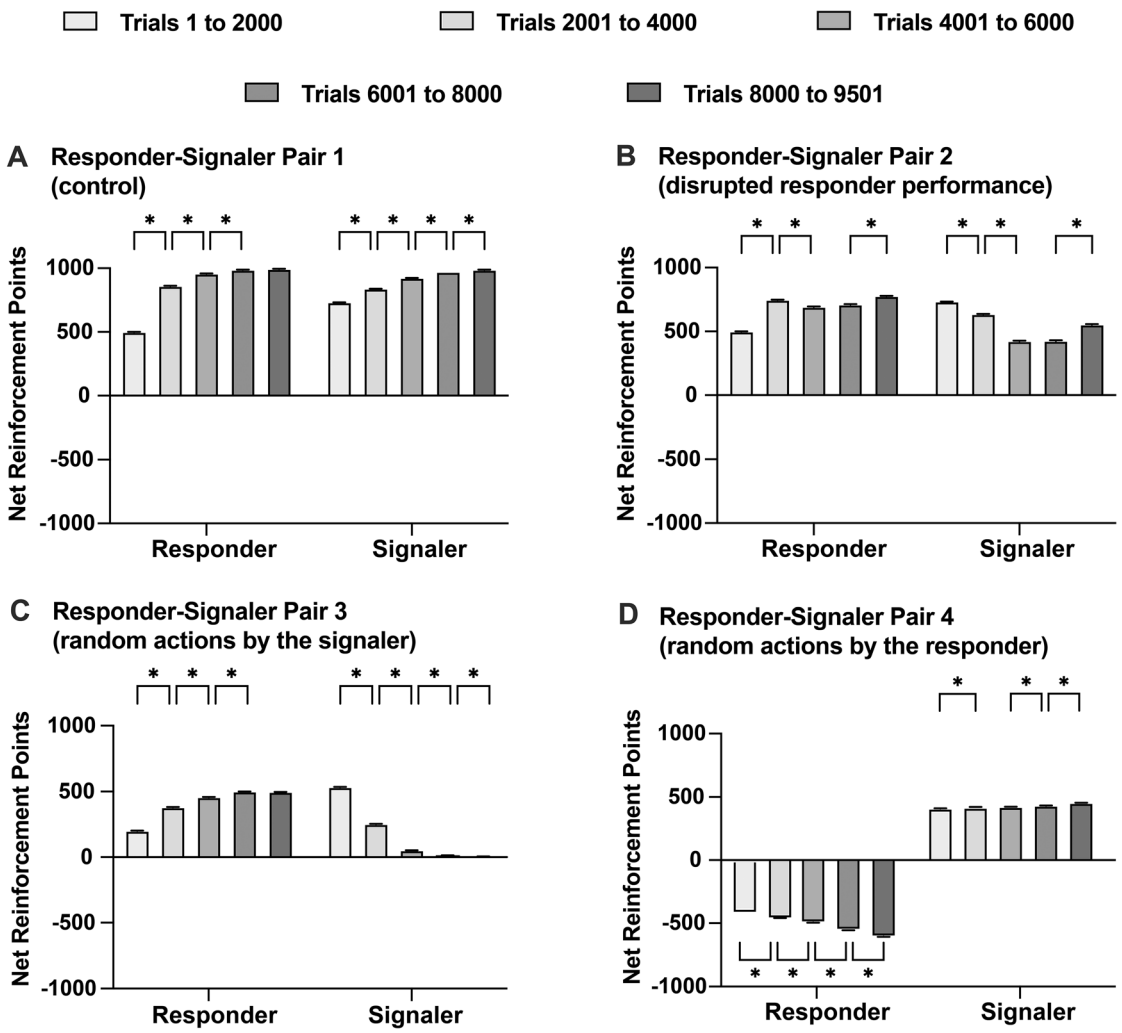

FIGURE 2 Moving averages of net reinforcement points of the virtual rat (Responder) and the virtual robot (Signaler) across trial bins (shades of gray) in different in silico pairs. Panel A: Responder–Signaler Pair 1 (control condition). Panel B: Responder–Signaler Pair 2 (disrupted responder performance). Panel C: Responder–Signaler Pair 3 (random action by the signaler). Panel D: Responder–Signaler Pair 4 (random action by the responder). Error bars depict clustered standard errors (*$p < .001$).

In Bin 3, the responders' performance followed the same trend as in Bin 2. The signaler in Pair 1 maintained peak performance, whereas those in Pairs 2, 3, and 4 performed worse. The signaler in Pair 2 still outperformed the signaler in Pair 3, and performance in Pairs 2 and 4 was similar, reflecting the effects of disrupted versus randomized manipulations. The signaler in Pair 4 performed better than the randomly behaving signaler in Pair 3 (Appendix F, Table F1).

In Bin 4, the performance patterns for both responders and signalers mirrored those of Bin 3 (Appendix E, Table E1; Appendix F, Table F1). In Bin 5, responder performance remained consistent with that in previous bins (Appendix E, Table E1). The signaler in Pair 1 (control) continued to outperform the others. However, unlike in Bin 4, the signaler in Pair 2 performed better than the one in Pair 4. The signaler in Pair 3, exhibiting random behavior, continued to see its reinforcement points drop to near zero (Appendix F, Table F1).

Discussion

Different in silico cases reflecting naturalistic scenarios tested how alterations in the behavior of one agent affected the performance of the other. The first virtual rat–robot pair, or Responder–Signaler Pair 1, simulating the unmanipulated or control scenario, revealed successful mutual learning, as both agents displayed reward maximization. This observation is in line with a recent human–robot mutual language-learning study, in which the learning performance of the human and the robot were positively correlated (Markelius et al., 2023). The reinforcement points of the responder converged in the last batch of trials, indicating a successful implementation of the Q-learning

algorithm (Baird, 1994). Although the signaler kept increasing its learning performance in the last batch of trials, it also reached a maximum reward level and displayed a flattened learning curve at the end of the simulation period (Mohr & van Rijn, 2022).

The responder in Pair 2, set to exhibit a low learning rate, increased its reinforcement points until the middle (Trial 4,000) of the simulation period, after which its performance fluctuated. This disruption corresponds to the demonstrated poor performance stemming from an initially low learning rate (Chen et al., 2021). The signaler displayed the same trend and exhibited poor performance, as shown before with a Nao robot and a low-performing human partner in a pattern recognition task (Kirtay et al., 2022). Randomized behavior of the signaler in Pair 3 initially increased the reinforcement points of the responder, which then converged on a suboptimal level in the last batch of trials. In contrast, the signaler in this pair initially displayed higher performance due to the early approaches made by the responder. The responder's suboptimal performance in this condition was in line with an earlier in vivo finding showing that randomized discriminative stimuli lead to suppressed learning compared with conditions that possess a rule-based order (Domenger & Schwarting, 2005). In the Responder–Signaler Pair 4, the randomly behaving responder displayed no learning and obtained accumulated punishments due to its signal-independent interaction efforts with the signaler throughout the trials. This led the signaler to receive some reinforcement points and display marginal increases in reinforcement points as observed before (Kirtay et al., 2022).

Mutual performance in the initial phase of trials was substantially reduced when one of the agents was set to behave randomly in Pair 3 and Pair 4. This suggests that although a low learning rate (Pair 2) leads to a fluctuated learning curve, it does not block mutual learning, as observed with randomized behavior of one of the partners. When both agents interacted with a random-behaving partner, the responder in Pair 3 displayed learning up to a certain point, whereas the signaler in Pair 4 did not show any learning. This difference may have emerged from the structure of the experimental paradigm, which gave the responder a more active role, as both the signaler's and the responder's reinforcement depended on the rat's approach behavior. The reinforcement of the signaler relied on responder's actions (i.e., entering the interaction zone). The reinforcement of the responder requires a positive signal from the signaler, but it is only partially affected by the learning rate of the signaler. Even if the signaler completely fails to acquire the rules of the paradigm, it starts the simulation with a 50% chance of presenting the positive signal, providing the rat with sufficient reward opportunity.

Overall, the in silico tutorial revealed Q-learning as a highly adaptable and balanced tool for piloting experiments that contain dynamic and interactive aspects. It demonstrated the feasibility of a mutual learning paradigm in which both parties simultaneously assumed both the role of a teacher and a student, teaching the other agent rules of the paradigm while receiving positive feedback to shape its own actions. Different paradigms that involve mutual interactions were tested with image segmentation (Zhang & Zhang, 2021) and visual question generation by using reinforcement learning (Xu et al., 2018). However, the present in silico experiments were the first to test a behavioral paradigm with two partners that independently move and act on each other. Once the feasibility of the mutual learning was demonstrated in a virtual setting, the paradigm was adapted to a real-world environment and tested with eight different rat–robot pairs.

IN VIVO EXPERIMENTS

Methods

Subjects

Adult female Wistar rats ($n = 8$; 152–180 g) were housed individually in the laboratory vivarium (21 ± 1°C with ~50% humidity; 12 hr light: 12 hr dark cycle, with lights on at 0700 hours). The animals were obtained from the Boğaziçi University Vivarium Colony (i.e., inbred), and only female rats were used due to availability. A 30-day food restriction regimen was initiated at the beginning of the habituation period and continued until the end of the experiment, during which animals received 5 g of diet per day. By the end of the restriction period, the average weight of the animals had decreased by 17% relative to their initial weight. Daily morning weigh-ins revealed that no animal's weight loss exceeded 19% of its starting weight at any point during the experiment. Experimental sessions were conducted between 0700 and 1100 hours each day. All procedures were approved by the Boğaziçi University Ethics Committee for the Use of Animals in Experiments.

Robot

The robot was built on a differential drive robotic platform (Seeed Studio Shield Bot). A microcontroller (Arduino Uno), an IMU module consisting of an accelerometer, a gyroscope, and a Wi-Fi module, were attached to the platform. The upper surface of the robot had a plastic cover to protect its circuit components from the rat. The body of the robot was designed in TinkerCAD and printed with a Prusa i3 MK3S+ 3D printer (Appendix G, Figure G1).

The robotic platform was powered by an internal battery. A control computer was used to send commands to the robot via Wi-Fi packets, and an internal microcontroller

module decoded and executed these command inputs. The computer relayed six high-level commands: (1) turn left, (2) turn right, (3) turn back from left, (4) turn back from right, (5) move straight to front, and (6) move straight to back. The microcontroller executed high-level commands using the current state readings from the IMU module. The accelerometer and gyroscope data were used to calculate the current attitude. The IMU module was automatically calibrated to minimize measurement errors at the beginning of each experiment.

In vivo trials and mutual learning

The robot used locomotor signals, turning 90° either to the right or left and moving 20 cm forward. This action served as a discriminative stimulus for reinforcement (S^{Dr}) or punishment (S^{Dp}) depending on the rat's response. The positive and negative signals were counterbalanced. In half of the rat–robot pairs, moving to the right denoted a positive signal and moving to the left indicated a negative signal across all days and trials of that pair. In the other half, moving to the left denoted a positive signal and moving to the right indicated a negative signal. After the signal, the robot turned back and moved to its starting point, the *robot zone* (diameter: 25 cm) located at the center of the arena. The complete movement of the robot from its 90° turn to arriving at its starting point and aligning itself to face the starting point of the rat took approximately 8 s, depending on the interaction window discussed below. It should be noted that these signals were accompanied by the motor sound of the robot when it was moving. This was a 50-dB noise as recorded from the starting point of the animal at the side of the arena.

Upon a positive signal, the rat had to move next to the robot to receive a reward, as in the virtual trials. The *interaction zone* (diameter: 25 cm) was located around the moving robot. As the robot relocated from its starting position, presenting a signal, the animal had to follow the robot within the arena. The time window during which the rat had to enter the interaction zone following a signal was designated as the *interaction window* (refer to Experimental schedule for the onset and offset of the interaction window). The rat had to approach the robot and enter the interaction zone within this time window to be reinforced for this behavior. Following interaction with the robot after a positive signal, a food reward would be dispensed from the feeder located in the top right-hand corner of the arena and the animal would then need to reach and consume the reward. The rat's behavior would be punished if it entered the interaction zone during the interaction window after a negative signal (refer to Figure 3 for the experimental setup and conditions).

The robot was rewarded for attracting the rat to itself. Its behavior was reinforced each time the rat entered the interaction zone following the robot's positive or negative signal. The robot's learning was achieved using the same rules as in the in silico experiments, with two important differences. First, a minimum epsilon value of 0.3 was imposed. This value was reached within 7 days of testing, as the robot started to favor exploitation over exploration. The epsilon value could be allowed get lower than 0.3 after this point, but the floor was maintained so that the robot had at least a 30% chance of behaving randomly. The second modification was done to the Q value, updating equation, which was changed to the following:

$$Q^{new}(s_t, a_t) \leftarrow \frac{Q(s_t, a_t) + \alpha \cdot r_t}{1 + \alpha}, \tag{6}$$

where $Q(s_t, a_t)$ is the current Q value of the action taken, α is the learning rate, and r_t is the reward as a result of the action.

These changes were implemented to accommodate the robot to the substantially low number of in vivo trials, as opposed to the 10,000-trial set of in silico experiments. The first change was intended to limit the convergence of the robot to a particular behavior. If the robot converged too fast, it would not be possible to understand whether the rat could differentiate between signals in the later trials of the experiment, as rapid converging would have led the robot to constantly give the same signal across trials. The second change was intended to allow the robot to rapidly update its Q values when the rat changed its behavior.

Apparatus

The experimental apparatus was a square (75 × 75 cm) arena made with black Plexiglas. A camera (HDR-CX240 Handycam, SONY, CA) was positioned 3 m above the center of the arena for video recording and online communication with the control computer. This computer functioned as the main interface unit between the rat and the robot, relaying real-time location of the animal via a DeepLabCut implementation described below.

A 3D-printed rotary feeder was attached to the top of the arena wall (75 cm) on the upper right corner to dispense the food reward. It contained a microcontroller (Arduino Uno), a servo motor, and a Wi-Fi module (ESP8266) to communicate with the control computer. The feeder held up 18 pieces of cocoa-flavored popped rice (Kellogg's Coco Pops/Cocoa Krispies) dropping one piece (3 g) toward the center of the arena as a reward.

The punishment in the in vivo experiments was a pure 1-kHz tone at 100 dB (refer to Friedel et al., 2017) that lasted as long as the rat was inside the interaction zone during the interaction window following a negative

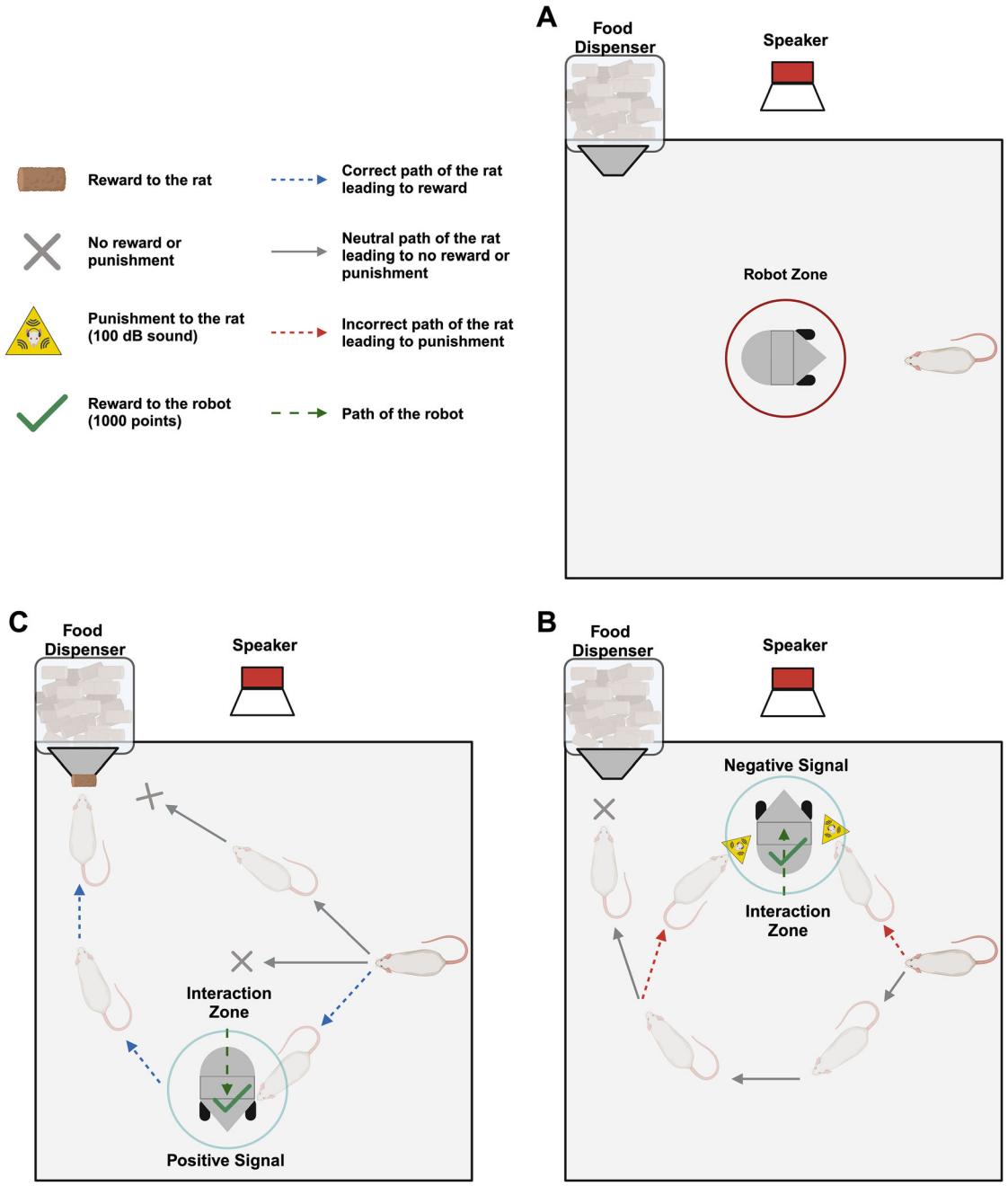

FIGURE 3 In vivo experimental setup and signal conditions. Panel A: The experimental setup is shown, with the robot and rat in their starting positions within the robot zone. Panel B: The positive signal condition is illustrated, where the rat's action is rewarded only if it interacts with the robot within the interaction zone and it is not rewarded if it enters the robot zone or the food dispenser. Panel C: The negative signal condition is shown, where a punishment is given if the rat interacts with the robot within the interaction zone and other behaviors receive neither a reward nor a punishment.

signal by the robot. Two speakers (Z333, Logitech) were used to deliver the loud punishment sound at 100 dB, as measured from the center of the experimental arena. The pure 1-kHz tone was generated using a sine wave generator (pySineWave) and measured for distortions using spectrogram software (Decibel X). The speakers were controlled by two synchronized computers in the test room.

DeepLabCut implementation

The real-time position of the rat in each trial was constantly relayed to the robot and the main script that controls the experiment via a DeepLabCut (Mathis et al., 2018) implementation. The DeepLabCut model was trained to infer the real-time position of the rat in the experimental apparatus (Isik & Unal, 2023). The training

set was prepared by using video footage recorded with the same camera. Five parts of the body were labeled (known as keypoints) to track the location of the animal and infer its body posture: the head, left and right ears, torso, and tail. The training set was augmented using the Img-Aug method of the DeepLabCut, and the model was trained on a NVidia 3090 GPU until no more improvement was observed in the inference error, referring to the difference between the output of the network and the human-labeled data points in the test data set.

The trained model was then exported to *DLC-Live* (Kane et al., 2020), a software library for using DeepLabCut models in real-time position inference and pose estimation. During the experiments, frames coming from the video camera were preprocessed and given to *DLC-Live*. *DLC-Live* outputs the position inferences of the keypoints along with a confidence value (0–1) for these inferences. Inferences below the confidence threshold (0.7) were discarded.

The OpenCV library (Bradski, 2000) was used to preprocess the frames. Each frame was cropped to only keep the experimental apparatus, and each frame was resized to maintain all the images at the same size across experiments. Preprocessing parameters were calibrated every day to counter the effects of small imperfections in the placement of the experimental setup or the video camera.

Communication between devices

The control computer ran a Python script that regulated the onset for all triggered events in the experiment, including when the robot would make a decision. The script ran the learning and decision algorithms of the robot and stored and updated its Q values. It used the position inference from *DLC-Live* to determine whether a rat–robot interaction (i.e., rat's entrance to the interaction zone) would lead to a punishment or a reinforcement, depending on the preceding signal.

The control computer and the Arduino devices communicated through Wi-Fi. The computer sent Wi-Fi packets that contained high-level commands for the robot (move left, move right, or return to starting position), the feeder (run/stop), and the computer controlling the speakers (on/off). The robot received information from the computer to turn 90° to the right or left and move 20 cm forward, signaling a reward or punishment in response to the rat's action. The robot turned to the robot zone after receiving a return command from the control computer. The feeder received the command to run and gave one piece of reward upon the rat's completion of a desired action. The microcontrollers of the robot and the feeder then executed the high-level commands themselves without further input from the control computer. The computers connected to the speakers received "on" or "off" commands from the control computer and delivered the aversive stimulus when necessary.

Manual control of the robot

The robot was designed to move autonomously without the involvement of the experimenter. It could tolerate a certain level of obstruction by the rat while moving and align itself to complete its course without hitting the animal. Rarely, rats climbed to the top of the robot while it was moving or strongly pushed and dislocated it, making autoalignment impossible. In these cases, the experimenter intervened to manually align the robot using a gamepad connected to a second computer in the test room. The robot ignored commands from the control computer when under manual control by the experimenter.

Experimental schedule

The in vivo experiments spanned 31 days, with the rat spending 20 min in the arena with the robot each day. The experimental process started with a gradual habituation period, which took 17 days. During Days 1–3, the rats were placed in the arena with a static robot that did not perform any action or produce any sound. On Days 4–6, the robot made a 10-s-long noise, similar to the noise it makes while moving, every 2 min. For the following 10 days, the robot moved 20 cm forward, paused for 5 s, and then returned to its original position, repeating this sequence every 2 min.

Mutual interaction started on Day 18 and lasted for 2 weeks. On Days 18–24, each rat–robot pair went through 11 trials. On Days 25–31, the intertrial interval was decreased to accommodate 15 trials per day. To facilitate the early phases of learning, the interaction window was set to 45 s during the first 3 days. As learning progressed, this duration was gradually reduced to 30 s on Day 21, 20 s on Day 23, 12 s on Day 26, and finally 7 s on Day 29.

Procedure

The robot was placed in the robot zone at the center of the arena and powered up. The rat was then brought to the test room and allowed to habituate in its cage for 10 min. The experimenter gently placed the rat to the side of the arena facing the front of the robot. The robot did not make any movement or noise during the first 2 min of each trial, allowing the rat to habituate to the new environment, and then presented its first signal by turning 90° and moving to the right or left. The interaction window timer started when the robot finished its move. The rat had to approach the robot and enter the interaction zone within this interval to receive a reward or a punishment contingent on its behavior depending on the robot's signal. If the robot gave the positive signal, the rat was rewarded immediately after it entered the interaction zone. If the robot gave the negative signal, the rat was immediately presented with the aversive auditory stimulus upon its entrance to the interaction zone. The punishment was

terminated when the rat left the interaction zone or the robot returned to the robot zone at the end of the interaction window. The robot was rewarded any time the rat entered the interaction zone during the interaction window, irrespective of the nature of the robot's preceding signal.

When the rat did not approach the robot during the interaction window, neither the rat's actions nor the robot's actions were rewarded or punished. The robot made its next decision at the end of the intertrial interval, initiating the next trial. Each experimental session lasted 20 min. It is important to note that, although the same robot hardware was used across all experiments, its learning process and decisions were independent for each rat–robot pair. Each pair represented a unique interaction, with both the rat and the robot beginning the experiment in a naïve state.

Results

We conducted experiments with eight different rat–robot pairs over the 2-week training/interaction period. Each day, we documented the rat's reward, punishment, and net reward, as well as the robot's reward. The rat's reward consisted of food given following correct behavior in a positive signal condition. In contrast, punishment referred to instances where the rat's incorrect behavior in response to a negative signal was followed by an aversive sound. The net reward for the rat was calculated by subtracting punishments from rewards each day. Regardless of signal type, the robot's reward included instances of interaction with the rat. Data analysis was performed using GraphPad Prism Version 9.5.1, with a significance level set at .05 for all tests. Tukey's HSD and Dunnett's tests were employed for post hoc multiple comparisons.

Overall learning performance was visualized by plotting the day-by-day cumulative data of the rat's reward, punishment, and net reward as well as the robot's reward for each rat–robot pair (Figure 4). The area under the curve (AUC) for each variable was employed to quantify the cumulative performance of each pair. The AUC values were determined using the trapezoid rule, involving drawing a line between individual data points and computing the area under the line. Although the baseline was set to 0 in the analysis, negative data points below the baseline were also considered by subtracting the area below the baseline from the area above. Figure 4 illustrates that the AUC values for all pairs across the four variables were nearly identical, except for Pair 6 and Pair 7, which did not exhibit mutual learning.

Comparison of different pairs

For the noncumulative performance analysis of rats and robots, a one-way analysis of variance (ANOVA) was conducted for each agent by aggregating data across days for each pair. Tukey's post hoc tests were used for pairwise comparisons (Appendix H, Table H1). The pairs differed in terms of the rat's reward, with rats in Pairs 6 and 7 obtaining significantly fewer rewards over 14 days than the rats in Pairs 1–5 and 8. No significant differences were observed between the rats in Pairs 6 and 7. Similarly, the rats in Pairs 1–5 and 8 performed equally well.

Differences were also found in terms of the rats' punishment. Similar to the reward findings, the rats in Pairs 6 and 7 were punished less frequently over 14 days than those in Pairs 1–5 and 8. No significant differences were observed between the rats in Pairs 6 and 7 or among the rats in Pairs 1–5 and 8. These differences were reflected in the net reward calculation: The rats in Pairs 6 and 7 scored lower than those in Pairs 1–5 and 8. Although the rat in Pair 4 scored higher than the rat in Pair 6, no significant differences were found between the rats in Pairs 4 and 7. Similarly, no differences were observed between the rats in Pairs 6 and 7.

The eight rat–robot pairs also differed in terms of the robot's reward. Robots in Pairs 6 and 7 earned fewer rewards than those in Pairs 1–5 and 8. However, no significant differences were found between the robots in Pairs 6 and 7, both of which failed to maximize their rewards. Similarly, no significant differences were observed among the successful robots in Pairs 1–5 and 8, suggesting consistency within these pairs.

Comparison of successful performance across days

To illustrate the learning curve of each agent, we analyzed performance changes over the 14-day interaction period by comparing daily performance to that of the first training day. For this analysis, we pooled the reward, punishment, and net reward data from the successful pairs that exhibited mutual learning (Pairs 1–5 and 8). There was no difference across the 14 days in terms of the reward obtained by rats, $F(3.80, 19.01) = 2.89$, $p = .052$, $\eta^2 = 0.37$ (degrees of freedom were corrected with Greenhouse–Geisser correction as $\varepsilon = 0.29$; repeated-measures ANOVA). However, a detailed comparison of daily performance with Day 1 using Dunnett's post hoc test revealed that rats exhibiting mutual learning earned significantly more rewards on Day 7 ($M = 9.17$, $SEM = 0.54$) than on Day 1 ($M = 5.50$, $SEM = 0.85$; $p < .05$; see Figure 5A).

The amount of daily punishment received did not vary across interaction days, $F(3.76, 18.79) = 1.38$, $p > .05$, $\eta^2 = 0.22$ (degrees of freedom were corrected with Greenhouse–Geisser correction as $\varepsilon = 0.29$; repeated-measures ANOVA). There was no difference between a particular day and Day 1 (Dunnett's post hoc test $p > .05$; Figure 5A). Similarly, the net reward of the rats did not differ between individual days,

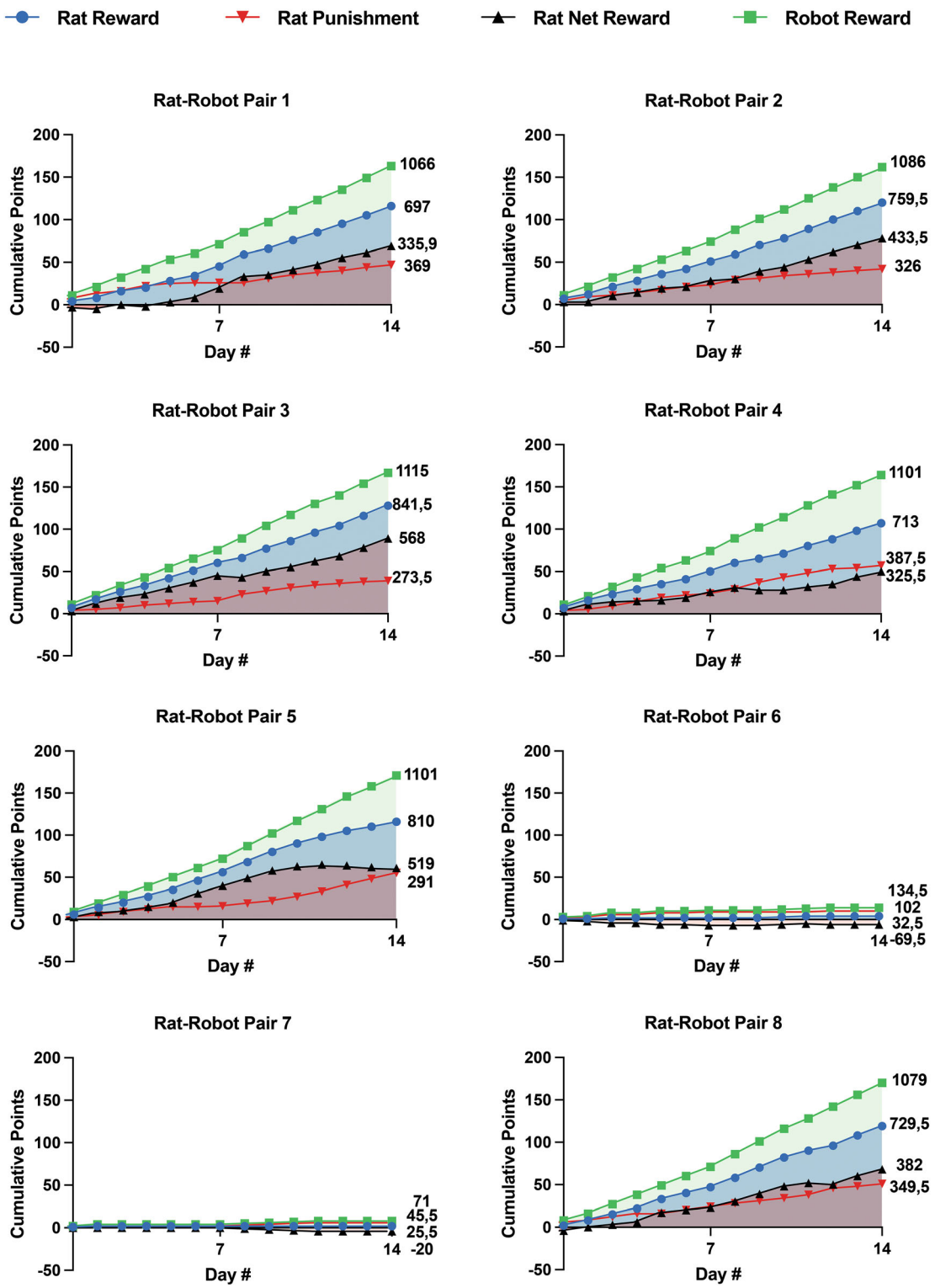

FIGURE 4 The cumulative points of reward (blue circles), punishment (red inverted triangles), net reward of the rat (black triangles), and reward of the robot (green squares) across interaction days in different in vivo pairs. The numbers on the right represent values for area under the curve.

$F(3.88, 19.37) = 1.31$, $p > .05$, $\eta^2 = 0.21$ (degrees of freedom were corrected with Greenhouse–Geisser correction as $\varepsilon = 0.30$; repeated-measures ANOVA). Yet, reflecting the difference observed for the reward, Dunnett's post hoc comparisons revealed that the animals obtained more net reward on Day 7 ($M = 7.50$, $SEM = 1.09$) than on Day 1 ($M = 0.83$, $SEM = 1.38$; $p < .05$; Figure 5A).

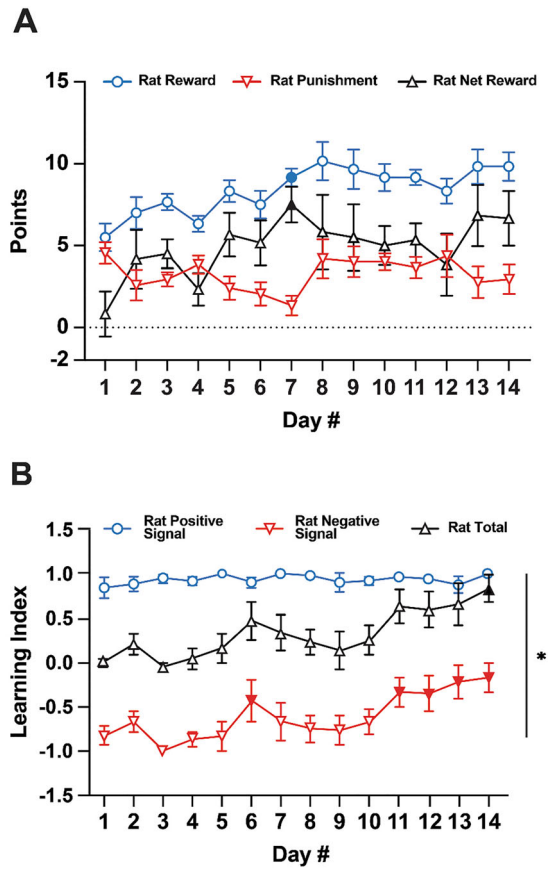

FIGURE 5 Averaged scores and learning indices of the rats in six in vivo pairs exhibiting mutual learning. Panel A: Mean values of reward (blue circles), punishment (red inverted triangles), and net reward (black triangles). Panel B: The positive signal (blue circles), negative signal (red inverted triangles), and total learning index (black triangles). Filled shapes indicate significant differences between the corresponding day and Day 1 ($p \leq .05$). Error bars depict SEM.

Finally, to evaluate the performance changes of the robots over 14 days, we calculated a proportion by dividing the robot's rewards on each day by the total number of trials: 11 for the first 7 days and 15 for the last 7 days. The total rewards obtained by the robots did not differ across the days, $F(3.13, 15.64) = 2.003$, $p > .05$, $\eta^2 = 0.03$ (degrees of freedom were corrected with the Greenhouse–Geisser correction as $\varepsilon = 0.24$; repeated-measures ANOVA).

Comparison of successful rat performance through learning indices

We calculated learning indices (−1 to +1) to effectively represent the learning performance of the successful rats (Pairs 1–5 and 8). Daily learning indices were derived for positive signal learning (Figure 5B, blue circles) and negative signal learning (Figure 5B, red inverted triangles), which were combined to obtain a total learning index (Figure 5B, black triangles). The positive signal learning index was calculated by subtracting the incorrect behavior (i.e., not interacting with the robot) following a positive signal from the correct behavior (i.e., interacting with the robot) and then dividing it by the number of positive signals presented by the robot. Likewise, the negative signal learning index was calculated by subtracting the incorrect behavior (i.e., interacting with the robot) following a negative signal from the correct behavior (i.e., not interacting with the robot) and then dividing it by the number of negative signals. The total learning index was calculated by summing the positive and negative signal learning indices and dividing the derived number by two.

We observed a main effect for the interaction day on the learning indices, $F(13, 65) = 2.68$, $p < .05$, $\omega^2 = 6.25$; two-way ANOVA). The positive signal learning index did not differ between Day 1 and subsequent days (Dunnett's test $p > .05$), but rats received a significantly higher score ($p < .05$) for the negative signal learning index on Day 6 ($M = -0.43$, $SEM = 0.24$), Day 11 ($M = -0.33$, $SEM = 0.17$), Day 12 ($M = -0.35$, $SEM = 0.21$), Day 13 ($M = -0.22$, $SEM = 0.20$), and Day 14 ($M = -0.17$, $SEM = 0.16$) than on Day 1 ($M = -0.83$, $SEM = 0.11$). Overall, in terms of the total learning index, rats that exhibited mutual learning displayed better performance ($p < .05$) on Day 14 ($M = 0.83$, $SEM = 0.16$) than on Day 1 ($M = 0.01$, $SEM = 0.05$).

There was a difference between different types of learning indices, $F(2, 10) = 1{,}488$, $p < .05$, $\omega^2 = 81.42$ (two-way ANOVA). Rats showed better performance under the positive signal ($M = 0.93$, $SEM = 0.01$) than under the negative signal ($M = -0.61$, $SEM = 0.07$; Tukey's HSD test $p < .05$). Their positive learning score ($M = 0.93$, $SEM = 0.01$) was higher than their total learning score ($M = 0.16$, $SEM = 0.04$; Tukey's HSD test $p < .05$), which was higher than the negative signal learning performance ($M = -0.61$, $SEM = 0.07$; Tukey's HSD test $p < .05$). Moreover, there was a significant interaction between test day and the type of learning index, $F(26, 130) = 2.11$, $p < .05$, $\omega^2 = 2.19$ (two-way ANOVA), indicating that learning to avoid punishment (i.e., negative signal learning index) is enhanced, whereas learning to seek reward (i.e., positive signal learning indices) remains stable for the rats and the robots (Figure 5).

Discussion

In the in vivo experiments, six pairs showed a steady increase in gaining rewards, whereas the remaining two showed no detectable change in their behavior, remaining apathetic toward each other throughout the 2-week interaction period. They did not engage with the robot under any signal condition. The indifferent behavior exhibited by the rats in these two pairs might be attributed to a lack of sufficient interest in the robot or a fear response

toward it. Rats have been shown to be more interested in biomimetic robots (Shi et al., 2011), and it is possible that the locomotor signals generated by the robot in this study were inadequate to capture the attention of the rats in these two pairs. Moreover, robots can evoke fear responses in rats (Shi, Ishii, Tanaka, et al., 2015), which may have contributed to the limited interaction between the partners in these two pairs. Importantly, these rats may have, by chance, approached the robot mostly following negative signals in the early phases of training, consequently facing punishment. This may have hindered the animals from pursuing further interaction with the robot, thereby restricting their ability to respond to positive signals in subsequent trials. In essence, the lack of success in these two pairs shows that mutual learning emerges when both parties exhibit comparable and proficient performance in the early phases of learning but that falters when one of the partners fails to cooperate.

Mutual learning was achieved when both partners in a pair equally mastered the rules of the paradigm, resulting in similar performance levels and reinforcement accumulation. This was observed in six rat–robot pairs, with mutual performance peaking around the middle of the 14-day interaction period. The rats were rewarded significantly more on Day 7, but the number of punishments remained stable. The observed trend aligns with previous research findings showing that rats can successfully acquire a discriminated operant behavior within a maximum of 7 days (Skelton et al., 1987) and mice display enhanced performance up to 7 days in a difficult stimulus-controlled operant conditioning task run by daily 20-min sessions (Hasan et al., 2013). Although the robots' rewards did not significantly vary across days—potentially due to periods when the robot was punished—the cumulative rewards showed a comparable increase, resulting in a similar learning curve for the robots. Thus, mutual learning occurs when both parties exhibit a comparable rate of learning (Markelius et al., 2023).

The mutual increase in the number of rewards of the partners was not reflected in the rats' punishments. Although successful animals reliably increased their reinforcement from the beginning of the trials until they peaked at the middle of the experimental period, the number of punishments they received following negative signals remained relatively stable throughout the trials. This difference likely arises due to the nature of the concept of punishment, which can be used to suppress unwanted behaviors in operant conditioning and reinforcement learning tasks (Bouton & Schepers, 2015). However, punishment is, by definition, not effective to teach a particular behavior to an organism or artificial agent, as it does not tell the details of the desired action but only signals that other behaviors are not wanted (Thorndike, 1927). Furthermore, the intensity of punishment is critical in changing behavior and the loud tone (1 kHz at 100 dB) used in the present study may have only affected two pairs and been insufficient for the animals that eventually exhibited mutual learning. Although the tone used in this study worked as a punishment (Friedel et al., 2017), similar auditory stimuli at 1 kHz are also used as conditioned stimuli and paired with a mild electric shock as a punishment in both in vivo (Boulanger Bertolus et al., 2015; Park et al., 2016; Rogan et al., 1997) and in silico studies (Rigoli et al., 2016).

The lack of a change in the number of punishments, however, does not indicate that the punishment used in this study had no effect on the behavior of the partners. It is important to note that the likelihood of receiving a punishment is linked to the negative signals from the robot and that the ratio of negative signals presented by the robot varies across trials as the robot adjusts its behavior. We have therefore calculated a more comprehensive metric, the learning indices, to account for the proportion of different signal types per interaction and be able to detect behavioral alterations that otherwise remain unnoticeable. Rats in the initial days of training immediately started to interact with the robot due to its novelty effect (Shi et al., 2011) and obtained a high number of rewards due to the interaction in the positive signal condition, albeit not yet learning the signals. This led to a relatively high, stable positive signal learning index throughout the experiments, as previously observed in Go/No-Go tasks in which the Go condition quickly elicits a high level of performance that remains stable as opposed to the No-Go behavior (Jones et al., 2017). Despite the relatively stable number of punishment points received by the rat's actions throughout the experiments, the negative signal learning index increased over time, reflecting a direct effect of punishment and indicating passive avoidance learning. As opposed to developing an active coping (Akmese et al., 2023), the rats learned not to interact with the robot under ambiguous signals to avoid punishment, as observed for other aversive stimuli like an air-puff (Moriarty et al., 2012) or footshock (Huang et al., 2013). Gradually increasing avoidance of the punishment suggests that the punishment led to a steady and permanent decrease in the unwanted behavior rather than a temporary suppression, as conceptualized in a recent study (Shahan et al., 2023).

GENERAL DISCUSSION

We tested four different scenarios in silico and observed that both virtual agents were able to adapt their behavior toward reward maximization and achieve mutual learning under control (unmanipulated) conditions. In vivo experiments tested this condition in eight different rat–robot pairs and led to successful mutual learning in six pairs. These rats rapidly acquired the behavior that was necessary to receive the reward and exhibited passive avoidance learning for negative signals. The robots in these pairs also displayed a steep learning curve and attracted the rats to themselves. When one of the partners failed to acquire the rules of the paradigm, the other also failed, as observed in two pairs.

The in silico experiments served as a tutorial for using the Q-learning algorithm in complex behavioral tasks

and demonstrated its utility in experiments involving multiple interacting agents. The algorithm's flexibility allowed for the piloting of a task involving adaptive behavior in a more naturalistic environment, enhancing our understanding of the mutual learning paradigm. This tutorial illustrates the algorithm's applicability to interaction paradigms involving both artificial and biological agents and highlights its potential for behavioral experiments, offering valuable insights for researchers studying mutual learning paradigms.

Classical mutual learning paradigms have been extensively studied in biological–biological interactions (Noë, 2006), where social cues, emotional aspects, and cooperation play significant roles. In contrast, biological–artificial interactions, such as those in this study, depend more on signaling reinforcements or punishments and maximizing rewards through interaction. Biological–biological interactions are characterized by a rich set of signals across multiple modalities, which facilitates learning (Laidre & Johnstone, 2013). In our study, the signaling strategy was much simpler, which might have limited the animals' ability to learn the rules of the paradigm.

The environment is also a crucial factor influencing biological–biological interactions. For instance, factors such as nest structure and weather have been shown to affect the personality of ant colonies (Pinter-Wollman et al., 2012). Although we sought to capture the complexity of behavior, the controlled environment and use of an artificial agent might have constrained the interactive learning process. Nevertheless, biological–artificial mutual learning paradigms offer the advantage of more controlled and detailed investigations into the dynamics of interactions, such as how the behavior of one agent influences the learning process over time. The use of robots in this context has significant implications, allowing the establishment of paradigms with multiple interacting agents, which would be challenging to achieve with other methods. As demonstrated in the in silico tutorial, the robot can be programmed to fit the intended paradigm and adapt its behavior in response to biological or artificial agents, making the interaction more akin to biological–biological interactions.

Future research could build on the foundation of our paradigm to develop more complex and sophisticated mutual learning scenarios. This could involve more intricate behavioral tasks, multiagent interactions, and varying environmental conditions. Adding such complexity could enhance our understanding of how different learning paradigms influence behavior and potentially lead to improved models and designs for human–robot interactions in educational and therapeutic settings.

This study serves as proof of concept, demonstrating the feasibility of bidirectional, or mutual, learning between a model organism and a reinforcement learning agent. Using in silico and in vivo experiments, we established an innovative rat–robot interaction paradigm that required cooperation between the partners to achieve their individual goals. The two agents interacted with each other to attain specific objectives, adapting their actions toward reward maximization based on signals from the partner, ultimately achieving mutual learning. When one of the partners failed to acquire the rules of the paradigm, either in silico or in vivo, the other also failed. The tested paradigm can be developed for use in other animal–machine interactions to test the efficacy of different learning rules and reinforcement schedules to achieve mutual learning between agents.

AUTHOR CONTRIBUTIONS

The first two authors (Nas and Albayrak) are co-first authors of the article with equal contribution. All authors contributed equally to the production of this article.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

ETHICS APPROVAL

All procedures were approved by the Boğaziçi University Ethics Committee for the Use of Animals in Experiments.

ORCID

Gunes Unal https://orcid.org/0000-0003-3013-0271

REFERENCES

Abbeel, P., & Ng, A. Y. (2004, July 4–8). *Apprenticeship learning via inverse reinforcement learning [Paper presentation]*. ICML '04: The Twenty-First International Conference on Machine learning, Banff, Alberta, Canada. https://doi.org/10.1145/1015330.1015430

Abdai, J., Korcsok, B., Korondi, P., & Miklósi, A. (2018). Methodological challenges of the use of robots in ethological research. Animal Behavior and *Cognition*, *5*(4), 326–340. 10.26451/abc.05.04.02.2018

Akmese, C., Sevinc, C., Halim, S., & Unal, G. (2023). Differential role of GABAergic and cholinergic ventral pallidal neurons in behavioral despair, conditioned fear memory and active coping. *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, *125*, Article 110760. https://doi.org/10.1016/J.PNPBP.2023.110760

Asadpour, M., Tâche, F., Caprari, G., Karlen, W., & Siegwart, R. (2006). Robot-animal interaction: Perception and behavior of insbot. *International Journal of Advanced Robotic Systems*, *3*(2), 93–98. https://doi.org/10.5772/5752

Baird, L. C. (1994). Reinforcement learning in continuous time: Advantage updating. *International Conference on Neural Networks*, *4*, 2448–2453. https://doi.org/10.1109/ICNN.1994.374604

Bierbach, D., Landgraf, T., Romanczuk, P., Lukas, J., Nguyen, H., Wolf, M., & Krause, J. (2018). Using a robotic fish to investigate individual differences in social responsiveness in the guppy. *Royal Society Open Science*, *5*(8), Article 181026. https://doi.org/10.1098/RSOS.181026

Boulanger Bertolus, J., Knippenberg, J., Verschueren, A., Le Blanc, P., Brown, B. L., Mouly, A. M., & Doyère, V. (2015). Temporal behavior in auditory fear conditioning: Stimulus property matters. *International Journal of Comparative Psychology*, *28*(1). 10.46867/IJCP.2015.28.02.04

Bouton, M. E., & Schepers, S. T. (2015). Renewal after the punishment of free operant behavior. *Journal of Experimental Psychology: Animal Learning and Cognition*, *41*(1), 81–90. https://doi.org/10.1037/XAN0000051

Bradski G. (2000). *Open CV* 2. [Computer software].

Buşoniu, L., Ernst, D., De Schutter, B., & Babuška, R. (2010). Approximate dynamic programming with a fuzzy parameterization. *Automatica*, *46*(5), 804–814. https://doi.org/10.1016/J.AUTOMATICA.2010.02.006

Cazenille, L., Collignon, B., Chemtob, Y., Bonnet, F., Gribovskiy, A., Mondada, F., Bredeche, N., & Halloy, J. (2018). How mimetic should a robotic fish be to socially integrate into zebrafish groups? *Bioinspiration & Biomimetics*, *13*(2), Article 025001. https://doi.org/10.1088/1748-3190/AA8F6A

Chemtob, Y., Cazenille, L., Bonnet, F., Gribovskiy, A., Mondada, F., & Halloy, J. (2020). Strategies to modulate zebrafish collective dynamics with a closed-loop biomimetic robotic system. *Bioinspiration & Biomimetics*, *15*(4), Article 046004. https://doi.org/10.1088/1748-3190/AB8706

Chen, Y., Schomaker, L., & Wiering, M. (2021, February 4–6). An investigation into the effect of the learning rate on overestimation bias of connectionist Q-learning. *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, *2*, 107–118. https://doi.org/10.5220/0010227301070118

Clifton, J., & Laber, E. (2020). Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, *7*, 279–301. https://doi.org/10.1146/ANNUREV-STATISTICS-031219-041220

Del Angel Ortiz, R., Contreras, C. M., Gutiérrez-Garcia, A. G., & González, M. F. M. (2016). Social interaction test between a rat and a robot: A pilot study. *International Journal of Advanced Robotic Systems*, *13*(1). https://doi.org/10.5772/62015

Domenger, D., & Schwarting, R. K. W. (2005). Sequential behavior in the rat: A new model using food-reinforced instrumental behavior. *Behavioural Brain Research*, *160*(2), 197–207. https://doi.org/10.1016/J.BBR.2004.12.002

Friedel, J. E., DeHart, W. B., & Odum, A. L. (2017). The effects of 100 dB 1-kHz and 22-kHz tones as punishers on lever pressing in rats. *Journal of the Experimental Analysis of Behavior*, *107*(3), 354–368. https://doi.org/10.1002/JEAB.254

Frohnwieser, A., Murray, J. C., Pike, T. W., & Wilkinson, A. (2016). Using robots to understand animal cognition. *Journal of the Experimental Analysis of Behavior*, *105*(1), 14–22. https://doi.org/10.1002/JEAB.193

Gianelli, S., Harland, B., & Fellous, J.-M. (2018). A new rat-compatible robotic framework for spatial navigation behavioral experiments. *Journal of Neuroscience Methods*, *294*, 40–50. https://doi.org/10.1016/j.jneumeth.2017.10.021

Graubard, B. I., & Korn, E. L. (1994). Regression analysis with clustered data. *Statistics in Medicine*, *13*(5–7), 509–522. https://doi.org/10.1002/SIM.4780130514

Hasan, M. T., Hernández-González, S., Dogbevia, G., Treviño, M., Bertocchi, I., Gruart, A., & Delgado-García, J. M. (2013). Role of motor cortex NMDA receptors in learning-dependent synaptic plasticity of behaving mice. *Nature Communications*, *4*(1), 1–10. https://doi.org/10.1038/ncomms3258

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*(2), 245–258. https://doi.org/10.1016/J.NEURON.2017.06.011

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*(4), 709–722. https://doi.org/10.3758/BF03192961

Huang, A. C. W., Shyu, B. C., Hsiao, S., Chen, T. C., & He, A. B. H. (2013). Neural substrates of fear conditioning, extinction, and spontaneous recovery in passive avoidance learning: A c-fos study in rats. *Behavioural Brain Research*, *237*(1), 23–31. https://doi.org/10.1016/J.BBR.2012.09.024

Ishii, H., Aoki, T., Moribe, K., Nakasuji, M., Miwa, H., & Takanishi, A. (2003, October 31–November 2). *Interactive experiments between creature and robot as a basic research for coexistence between human and robot [Paper presentation]*. 12th IEEE International Workshop on Robot and Human Interactive Communication, Millbrae, CA, USA. https://doi.org/10.1109/ROMAN.2003.1251870

Ishii, H., Ogura, M., Kurisu, S., Komura, A., Takanishi, A., Iida, N., & Kimura, H. (2006, September 25–29). *Experimental study on task teaching to real rats through interaction with a robotic rat [Paper presentation]*. 9th International Conference on Simulation of Adaptive Behavior, SAB 2006, Rome, Italy. https://doi.org/10.1007/11840541_53

Isik, S., & Unal, G. (2023). Open-source software for automated rodent behavioral analysis. *Frontiers in Neuroscience*, *17*, Article 1149027. https://doi.org/10.3389/FNINS.2023.1149027

Jones, S., Paul, E. S., Dayan, P., Robinson, E. S. J., & Mendl, M. (2017). Pavlovian influences on learning differ between rats and mice in a counter-balanced Go/NoGo judgement bias task. *Behavioural Brain Research*, *331*, 214–224. https://doi.org/10.1016/J.BBR.2017.05.044

Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A., & Mathis, M. W. (2020). Real-time, low-latency closed-loop feedback using markerless posture tracking. *ELife*, *9*, 1–29. https://doi.org/10.7554/ELIFE.61909

Kirtay, M., Oztop, E., Kuhlen, A. K., Asada, M., & Hafner, V. V. (2022, August 29–September 2). *Trustworthiness assessment in multimodal human-robot interaction based on cognitive load [Paper presentation]*. 31st IEEE International Conference on Robot and Human Interactive Communication: Social, Asocial, and Antisocial Robots, Naples, Italy. https://doi.org/10.1109/RO-MAN53752.2022.9900730

Klein, B. A., Stein, J., & Taylor, R. C. (2012). Robots in the service of animal behavior. *Communicative & Integrative Biology*, *5*(5), 466–472. https://doi.org/10.4161/CIB.21304

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *International Journal of Robotics Research*, *32*(11), 1238–1274. https://doi.org/10.1177/0278364913495721

Krause, J., Winfield, A. F. T., & Deneubourg, J. L. (2011). Interactive robots in experimental biology. *Trends in Ecology & Evolution*, *26*(7), 369–375. https://doi.org/10.1016/J.TREE.2011.03.015

Laidre, M. E., & Johnstone, R. A. (2013). Animal signals. *Current Biology*, *23*, R829–R833.

Li, L., Ravi, S., & Wang, C. (2022). Editorial: Robotics to understand animal behaviour. *Frontiers in Robotics and AI*, *9*, Article 963416. https://doi.org/10.3389/FROBT.2022.963416

Ludvig, E. A., Bellemare, M. G., & Pearson, K. G. (2010). A primer on reinforcement learning in the brain: Psychological, computational, and neural perspectives. In E. Alonso & E. Mondragón (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications* (pp. 111–144). https://doi.org/10.4018/978-1-60960-021-1.CH006

Markelius, A., Sjöberg, S., Lemhauori, Z., Cohen, L., Bergström, M., Lowe, R., & Cañamero, L. (2023). A human-robot mutual learning system with affect-grounded language acquisition and differential outcomes training. In A. A. Abdulaziz, John-John Cabibihan, Nader Meskin, Silvia Rossi, Wanyue Jiang, Hongsheng He, & Shuzhi Sam Ge (Eds.), *Social Robotics* (pp. 108–122). Springer. https://doi.org/10.1007/978-981-99-8718-4_10

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289. https://doi.org/10.1038/s41593-018-0209-y

Miklósi, Á., & Gerencsér, L. (2012, December 2–5). *Potential application of autonomous and semi-autonomous robots in the study of animal behaviour [Paper presentation]*. 3rd IEEE International Conference on Cognitive Infocommunications, Košice, Slovakia. https://doi.org/10.1109/COGINFOCOM.2012.6421952

Mohr, F., & van Rijn, J. N. (2022). *Learning curves for decision making in supervised machine learning: A Survey*. arXiv, Cornell University. 10.48550/arXiv.2201.12150

Moriarty, O., Roche, M., McGuire, B. E., & Finn, D. P. (2012). Validation of an air-puff passive-avoidance paradigm for assessment of

aversive learning and memory in rat models of chronic pain. *Journal of Neuroscience Methods*, *204*(1), 1–8. https://doi.org/10.1016/J.JNEUMETH.2011.10.024

Noë, R. (2006). Cooperation experiments: coordination through communication versus acting apart together. *Animal Behaviour*, *71*, 1–18. https://doi.org/10.1016/j.anbehav.2005.03.037

Park, S., Lee, J., Park, K., Kim, J., Song, B., Hong, I., Kim, J., Lee, S., & Choi, S. (2016). Sound tuning of amygdala plasticity in auditory fear conditioning. *Scientific Reports*, *6*(1), 1–14. https://doi.org/10.1038/srep31069

Peng, X. Bin, Coumans, E., Zhang, T., Lee, T. W. E., Tan, J., & Levine, S. (2020, July 12–16). *Learning agile robotic locomotion skills by imitating animals [Paper presentation]*. Robotics: Science and Systems, Corvalis, Oregon, USA. 10.15607/RSS.2020.XVI.064

Pinter-Wollman, N., Gordon, D. M., & Holmes, S. (2012). Nest site and weather affect the personality of harvester ant colonies. *Behavioral Ecology*, *23*(5), 1022–1029. https://doi.org/10.1093/beheco/ars066

Quinn, L. K., Schuster, L. P., Aguilar-Rivera, M., Arnold, J., Ball, D., Gygi, E., Heath, S., Holt, J., Lee, D. J., Taufatofua, J., Wiles, J., & Chiba, A. A. (2018). When rats rescue robots. *Animal Behavior and Cognition*, *5*(4), 368–379. 10.26451/abc.05.04.04.2018

Rigoli, F., Pezzulo, G., & Dolan, R. J. (2016). Prospective and Pavlovian mechanisms in aversive behaviour. *Cognition*, *146*, 415–425. https://doi.org/10.1016/J.COGNITION.2015.10.017

Rogan, M. T., Staubli, U. V., & LeDoux, J. E. (1997). Fear conditioning induces associative long-term potentiation in the amygdala. *Nature*, *390*, 604–607. https://doi.org/10.1038/37601

Romano, D., Donati, E., Benelli, G., & Stefanini, C. (2018). A review on animal–robot interaction: From bio-hybrid organisms to mixed societies. *Biological Cybernetics*, *113*(3), 201–225. https://doi.org/10.1007/S00422-018-0787-5

Rundus, A. S., Owings, D. H., Joshi, S. S., Chinn, E., & Giannini, N. (2007). Ground squirrels use an infrared signal to deter rattlesnake predation. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(36), 14372–14376. https://doi.org/10.1073/PNAS.0702599104

Shahan, T. A., Sutton, G. M., Nist, A. N., & Davison, M. (2023). Aversive control versus stimulus control by punishment. *Journal of the Experimental Analysis of Behavior*, *119*(1), 104–116. https://doi.org/10.1002/JEAB.805

Shi, Q., Ishii, H., Fumino, S., Konno, S., Kinoshita, S., Takanishi, A., Okabayashi, S., Iida, N., & Kimura, H. (2011, December 7–11). *A robot-rat interaction experimental system based on the rat-inspired mobile robot WR-4 [Paper presentation]*. 2011 IEEE International Conference on Robotics and Biomimetics, ROBIO 2011, Karon Beach, Thailand. https://doi.org/10.1109/ROBIO.2011.6181319

Shi, Q., Ishii, H., Kinoshita, S., Konno, S., Takanishi, A., Okabayashi, S., Iida, N., & Kimura, H. (2013). A rat-like robot for interacting with real rats. *Robotica*, *31*(8), 1337–1350. https://doi.org/10.1017/S0263574713000568

Shi, Q., Ishii, H., Kinoshita, S., Takanishi, A., Okabayashi, S., Iida, N., Kimura, H., & Shibata, S. (2013). Modulation of rat behaviour by using a rat-like robot. *Bioinspiration & Biomimetics*, *8*(4), Article 046002. https://doi.org/10.1088/1748-3182/8/4/046002

Shi, Q., Ishii, H., Sugahara, Y., Takanishi, A., Huang, Q., & Fukuda, T. (2015). Design and control of a biomimetic robotic rat for interaction with laboratory rats. *IEEE/ASME Transactions on Mechatronics*, *20*(4), 1832–1842. https://doi.org/10.1109/TMECH.2014.2356595

Shi, Q., Ishii, H., Tanaka, K., Sugahara, Y., Takanishi, A., Okabayashi, S., Huang, Q., & Fukuda, T. (2015). Behavior modulation of rats to a robotic rat in multi-rat interaction. *Bioinspiration & Biomimetics*, *10*(5), Article 056011. https://doi.org/10.1088/1748-3190/10/5/056011

Silveira, P. S. P., de Oliveira Siqueira, J., Bernardy, J. L., Santiago, J., Meneses, T. C., Portela, B. S., & Benvenuti, M. F. (2023). Modeling VI and VDRL feedback functions: Searching normative rules through computational simulation. *Journal of the Experimental Analysis of Behavior*, *119*(2), 324–336. https://doi.org/10.1002/JEAB.826

Skelton, R. W., Scarth, A. S., Wilkie, D. M., Miller, J. J., & Phillips, A. G. (1987). Long-term increases in dentate granule cell responsivity accompany operant conditioning. *Journal of Neuroscience*, *7*(10), 3081–3087. https://doi.org/10.1523/JNEUROSCI.07-10-03081.1987

Skinner, B. F. (1932). On the rate of formation of a conditioned reflex. *Journal of General Psychology*, *7*(2), 274–286. https://doi.org/10.1080/00221309.1932.9918467

Skinner, B. F. (1935). The generic nature of the concepts of stimulus and response. *The Journal of General Psychology*, *12*(1), 40–65. https://doi.org/10.1080/00221309.1935.9920087

Son, J. H., Choi, Y. C., & Ahn, H. S. (2014). Bio-insect and artificial robot interaction using cooperative reinforcement learning. *Applied Soft Computing*, *25*, 322–335. https://doi.org/10.1016/J.ASOC.2014.09.002

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). The MIT Press.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), i–109. https://doi.org/10.1037/H0092987

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222. https://doi.org/10.2307/1415413

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* [Doctoral dissertation, King's College]. https://www.cs.rhul.ac.uk/∼chrisw/new_thesis.pdf

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3–4), 279–292. https://doi.org/10.1007/BF00992698

Xie, H., Gao, Z., Jia, G., Shimoda, S., & Shi, Q. (2023). Learning rat-like behavioral interaction using a small-scale robotic rat. *Cyborg and Bionic Systems*, *4*, Article 0032. 10.34133/CBSYSTEMS.0032

Xie, H., Jia, G., Al-Khulaqui, M., Gao, Z., Guo, X., Fukuda, T., & Shi, Q. (2022). A motion generation strategy of robotic rat using imitation learning for behavioral interaction. *IEEE Robotics and Automation Letters*, *7*(3), 7351–7358. https://doi.org/10.1109/LRA.2022.3182472

Xu, X., Song, J., Lu, H., He, L., Yang, Y., & Shen, F. (2018, July 23–27). *Dual learning for visual question generation [Paper presentation]*. 2018 IEEE International Conference on Multimedia and Expo, San Diego, CA, USA. https://doi.org/10.1109/ICME.2018.8486475

Zawadzki, E., Lipson, A., & Leyton-Brown, K. (2014). *Empirically evaluating multiagent learning algorithms*. arXiv. 10.48550/arXiv.1401.8074

Zhang, Y., & Zhang, J. (2021, October 29–November 1). *Dual-task mutual learning for semi-supervised medical image segmentation [Paper presentation]*. 4th Chinese Conference on Pattern Recognition and Computer Vision, Beijing, China. https://doi.org/10.1007/978-3-030-88010-1_46

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nas, O., Albayrak, D., & Unal, G. (2025). Of rats and robots: A mutual learning paradigm. *Journal of the Experimental Analysis of Behavior*, *123*(2), 176–201. https://doi.org/10.1002/jeab.70004

APPENDIX A

TABLE A1 Clustered regression summary for responder performance in different pairs.

Bin number	ß	CSE	*t*	*p*	Fit
Responder–Signaler Pair 1 (control)					
1 (intercept)	499.7836	1.674463	630.0	<.001	
2	361.3453	1.723665	322.1	<.001	
3	457.6775	1.679170	407.9	<.001	
4	487.1182	1.675204	434.2	<.001	
5	494.9597	1.674744	408.5	<.001	
R^2_{adj}					.779
Responder–Signaler Pair 2 (disrupted rat performance)					
1 (intercept)	499.322	1.829200	293.74	<.001	
2	248.430	2.127451	103.34	<.001	
3	193.441	2.593919	80.47	<.001	
4	211.968	2.691590	88.17	<.001	
5	277.485	2.598881	106.89	<.001	
R^2_{adj}					.171
Responder–Signaler Pair 3 (random actions by the robot)					
1 (intercept)	201.4349	0.8032037	452.8	<.001	
2	179.6258	0.9146777	285.5	<.001	
3	256.6050	0.8396717	407.9	<.001	
4	298.5934	0.8189223	474.6	<.001	
5	295.5983	0.8228196	435.1	<.001	
R^2_{adj}					.800
Responder–Signaler Pair 4 (random actions by the rat)					
1 (intercept)	−397.4830	0.3542944	−683.38	<.001	
2	−59.0076	0.5665424	−71.73	<.001	
3	−95.7671	0.6250394	−116.42	<.001	
4	−153.0236	0.7936390	−186.03	<.001	
5	−207.0708	1.0212109	−233.11	<.001	
R^2_{adj}					.473

Note: CSE = clustered standard error.

APPENDIX B

TABLE B1 Clustered regression summary for signaler performance in different pairs.

Bin number	ß	CSE	*t*	*p*	Fit
Responder–Signaler Pair 1 (control)					
1 (intercept)	731.7955	0.409442	2,820.69	<.001	
2	106.7205	0.503641	290.87	<.001	
3	191.9915	0.461257	523.28	<.001	
4	238.7600	0.431110	650.75	<.001	
5	256.5050	0.415399	647.37	<.001	
R^2_{adj}					.892
Responder–Signaler Pair 2 (disrupted rat performance)					
1 (intercept)	733.5870	0.462149	253.85	<.001	
2	−98.1398	2.100678	−24.01	<.001	
3	−309.7355	3.521894	−75.79	<.001	
4	−306.9610	3.952191	−75.11	<.001	
5	−179.3368	3.716996	−40.63	<.001	
R^2_{adj}					.102
Responder–Signaler Pair 3 (random actions by the robot)					
1 (intercept)	534.4945	0.760310	1,004.35	<.001	
2	−282.0765	0.977300	- 374.80	<.001	
3	- 482.3538	0.893196	- 640.91	<.001	
4	- 520.7057	0.822103	- 691.86	<.001	
5	- 539.3736	0.841002	- 663.63	<.001	
R^2_{adj}					.903
Responder–Signaler Pair 4 (random actions by the rat)					
1 (intercept)	408.0720	0.335833	1,184.68	<.001	
2	10.5588	0.458001	21.68	<.001	
3	11.9457	0.504069	24.52	<.001	
4	22.2882	0.454448	45.75	<.001	
5	44.8823	0.569025	85.32	<.001	
R^2_{adj}					.107

Note: CSE = clustered standard error.

APPENDIX C

TABLE C1 Responder performance: Comparing within responder–signaler pairs.

	First bin		Second bin					
	M	CSE	*M*	CSE	*t*	df	*p*	*d*
Responder–Signaler Pair 1 (control)								
Bin 1 vs. Bin 2	499.78	1.67	861.13	1.72	209.64	7	<.001	3.31
Bin 2 vs. Bin 3	861.13	1.72	957.46	1.68	40.03	7	<.001	0.63
Bin 3 vs. Bin 4	957.46	1.68	986.90	1.68	12.41	7	<.001	0.20
Bin 4 vs. Bin 5	986.90	1.68	994.74	1.67	3.31	7	>.001	0.06
Responder–Signaler Pair 2 (disrupted rat performance)								
Bin 1 vs. Bin 2	499.32	1.83	747.75	2.13	116.77	7	<.001	1.85
Bin 2 vs. Bin 3	747.75	2.13	692. 76	2.59	−16.39	7	<.001	−0.26
Bin 3 vs. Bin 4	692. 76	2.59	711.29	2.69	4.96	7	>.001	0.08
Bin 4 vs. Bin 5	711.29	2.69	776.81	2.60	17.51	7	<.001	0.30
Responder–Signaler Pair 3 (random actions by the robot)								
Bin 1 vs. Bin 2	201.43	0.80	381.06	0.91	196.38	7	<.001	3.11
Bin 2 vs. Bin 3	381.06	0.91	458.04	0.84	62.00	7	<.001	0.98
Bin 3 vs. Bin 4	458.04	0.84	500.02	0.82	35.80	7	<.001	0.57
Bin 4 vs. Bin 5	500.02	0.82	497.03	0.82	− 2.58	7	>.001	−0.04
Responder–Signaler Pair 4 (random actions by the rat)								
Bin 1 vs. Bin 2	−397.48	0.35	−456.49	0.57	−104.15	7	<.001	−1.65
Bin 2 vs. Bin 3	−456.49	0.57	−552.26	0.63	−43.58	7	<.001	−0.69
Bin 3 vs. Bin 4	−552.26	0.63	−705.28	0.79	−56.68	7	<.001	−0.90
Bin 4 vs. Bin 5	−705.28	0.79	−912.35	1.02	−41.79	7	<.001	−0.71

Note: *M* = mean; CSE = clustered standard error; df = degrees of freedom.

APPENDIX D

TABLE D1 Signaler performance: Comparing within responder–signaler pairs.

	First bin		Second bin					
	M	CSE	*M*	CSE	*t*	df	*p*	*d*
Responder–Signaler Pair 1 (control)								
Bin 1 vs. Bin 2	865.90	0.20	919.26	0.25	211.90	7	<.001	3.35
Bin 2 vs. Bin 3	919.26	0.25	961.89	0.23	124.85	7	<.001	1.97
Bin 3 vs. Bin 4	961.89	0.23	985.28	0.22	74.08	7	<.001	1.17
Bin 4 vs. Bin 5	985.28	0.22	994.15	0.21	29.64	7	<.001	0.50
Responder–Signaler Pair 2 (disrupted rat performance)								
Bin 1 vs. Bin 2	733.59	0.46	635.45	2.10	−46.72	7	<.001	−0.74
Bin 2 vs. Bin 3	635.45	2.10	423.83	3.52	−51.60	7	<.001	−0.82
Bin 3 vs. Bin 4	423.83	3.52	426.63	3.95	0.52	7	>.001	0.01
Bin 4 vs. Bin 5	426.63	3.95	554.25	3.72	23.52	7	<.001	0.40
Responder–Signaler Pair 3 (random actions by the robot)								
Bin 1 vs. Bin 2	534.49	0.76	252.41	0.98	−288.63	7	<.001	−4.56
Bin 2 vs. Bin 3	252.41	0.98	52.14	0.89	−151.270	7	<.001	−2.39
Bin 3 vs. Bin 4	52.14	0.89	13.78	0.82	−31. 59	7	<.001	−0.50
Bin 4 vs. Bin 5	13.78	0.82	−4.88	0.84	−15.87	7	<.001	−0.27
Responder–Signaler Pair 4 (random actions by the rat)								
Bin 1 vs. Bin 2	408.07	0.34	418.63	0.46	23.06	7	<.001	0.36
Bin 2 vs. Bin 3	418.63	0.46	420.01	0.51	2.04	7	>.001	0.03
Bin 3 vs. Bin 4	430.58	0.51	452.86	0.45	15.24	7	<.001	0.24
Bin 4 vs. Bin 5	452.86	0.45	497.75	0.57	31.03	7	<.001	0.52

Note: *M* = mean; CSE = clustered standard error; df = degrees of freedom.

APPENDIX E

TABLE E1 Responder performance: Comparing between responder–signaler pairs.

	First bin		Second bin					
	M	CSE	*M*	CSE	*t*	df	*p*	*d*
Bin 1								
Pair 1 vs. Pair 2	499.78	1.67	499.32	1.83	0.19	15	> 001	0.002
Pair 1 vs. Pair 3	499.78	1.67	201.43	0.80	160.65	15	<.001	2.54
Pair 1 vs. Pair 4	499.78	1.67	−397.48	0.35	524.25	15	<.001	8.29
Pair 2 vs. Pair 3	499.32	1.83	201.43	0.80	149.11	15	<.001	2.36
Pair 2 vs. Pair 4	499.32	1.83	−397.48	0.35	481.33	15	<.001	7.61
Pair 3 vs. Pair 4	201.43	0.80	−397.48	0.35	682.24	15	<.001	10.79
Bin 2								
Pair 1 vs. Pair 2	861.13	1.72	747.75	2.13	30.69	15	<.001	0.49
Pair 1 vs. Pair 3	861.13	1.72	381.06	0.91	178.21	15	<.001	2.81
Pair 1 vs. Pair 4	861.13	1.72	−456.49	0.57	528.26	15	<.001	8.35
Pair 2 vs. Pair 3	747.75	2.13	381.06	0.91	119.90	15	<.001	1.90
Pair 2 vs. Pair 4	747.75	2.13	−456.49	0.57	417.53	15	<.001	6.60
Pair 3 vs. Pair 4	381.06	0.91	−456.49	0.57	603.15	15	<.001	9.54
Bin 3								
Pair 1 vs. Pair 2	957.46	1.68	692. 76	2.59	52.82	15	<.001	0.84
Pair 1 vs. Pair 3	957.46	1.68	458.04	0.84	140.16	15	<.001	2.21
Pair 1 vs. Pair 4	957.46	1.68	−552.26	0.63	430.67	15	<.001	6.80
Pair 2 vs. Pair 3	692. 76	2.59	458.04	0.84	44.75	15	<.001	0.71
Pair 2 vs. Pair 4	692. 76	2.59	−552.26	0.63	231.87	15	<.001	3.67
Pair 3 vs. Pair 4	458.04	0.84	−552.26	0.63	557.57	15	<.001	8.82
Bin 4								
Pair 1 vs. Pair 2	986.90	1.68	711.29	2.69	47.80	15	<.001	0.76
Pair 1 vs. Pair 3	986.90	1.68	500.02	0.82	128.94	15	<.001	2.04
Pair 1 vs. Pair 4	986.90	1.68	−705.28	0.79	428.58	15	<.001	6.77
Pair 2 vs. Pair 3	711.29	2.69	500.02	0.82	42.51	15	<.001	0.67
Pair 2 vs. Pair 4	711.29	2.69	−705.28	0.79	261.34	15	<.001	4.13
Pair 3 vs. Pair 4	500.02	0.82	−705.28	0.79	505.18	15	<.001	7.98
Bin 5								
Pair 1 vs. Pair 2	994.74	1.67	776.81	2.60	33.31	15	<.001	0.61
Pair 1 vs. Pair 3	994.74	1.67	497.03	0.82	118.17	15	<.001	2.16
Pair 1 vs. Pair 4	994.74	1.67	−912.35	1.02	391.17	15	<.001	7.14
Pair 2 vs. Pair 3	776.81	2.60	497.03	0.82	49.35	15	<.001	0.90
Pair 2 vs. Pair 4	776.81	2.60	−912.35	1.02	247.66	15	<.001	4.52
Pair 3 vs. Pair 4	497.03	0.82	−912.35	1.02	448.07	15	<.001	8.18

Note: *M* = mean; CSE = clustered standard error; df = degrees of freedom.

APPENDIX F

TABLE F1 Signaler performance: Comparing between responder–signaler pairs.

	First bin		Second bin					
	M	*CSE*	*M*	*CSE*	*t*	df	*p*	*d*
Bin 1								
Pair 1 vs. Pair 2	865.90	0.20	733.59	0.46	−2.90	15	>.001	−0.04
Pair 1 vs. Pair 3	865.90	0.20	534.49	0.76	228.48	15	<.001	3.61
Pair 1 vs. Pair 4	865.90	0.20	408.07	0.34	611.31	15	<.001	9.66
Pair 2 vs. Pair 3	733.59	0.46	534.49	0.76	223.76	15	<.001	3.54
Pair 2 vs. Pair 4	733.59	0.46	408.07	0.34	569.80	15	<.001	9.01
Pair 3 vs. Pair 4	534.49	0.76	408.07	0.34	152.101	15	<.001	2.41
Bin 2								
Pair 1 vs. Pair 2	919.26	0.25	635.45	2.10	90.38	15	<.001	1.43
Pair 1 vs. Pair 3	919.26	0.25	252.41	0.98	419.23	15	<.001	6.62
Pair 1 vs. Pair 4	919.26	0.25	418.63	0.46	486.84	15	<.001	7.70
Pair 2 vs. Pair 3	635.45	2.10	252.41	0.98	154.33	15	<.001	2.44
Pair 2 vs. Pair 4	635.45	2.10	418.63	0.46	97.46	15	<.001	1.54
Pair 3 vs. Pair 4	252.41	0.98	418.63	0.46	−122.01	15	<.001	−1.93
Bin 3								
Pair 1 vs. Pair 2	961.89	0.23	423.83	3.52	119.34	15	<.001	1.89
Pair 1 vs. Pair 3	961.89	0.23	52.14	0.89	516.28	15	<.001	8.16
Pair 1 vs. Pair 4	961.89	0.23	420.01	0.51	481.19	15	<.001	7.61
Pair 2 vs. Pair 3	423.83	3.52	52.14	0.89	52.23	15	<.001	0.83
Pair 2 vs. Pair 4	423.83	3.52	420.01	0.51	0.55	15	>.001	0.01
Pair 3 vs. Pair 4	52.14	0.89	420.01	0.51	−218.97	15	<.001	−3.46
Bin 4								
Pair 1 vs. Pair 2	985.28	0.22	426.63	3.95	94.02	15	<.001	1.48
Pair 1 vs. Pair 3	985.28	0.22	13.78	0.82	489.09	15	<.001	7.73
Pair 1 vs. Pair 4	985.28	0.22	452.86	0.45	426.65	15	<.001	6.74
Pair 2 vs. Pair 3	426.63	3.95	13.78	0.82	69.14	15	<.001	1.09
Pair 2 vs. Pair 4	426.63	3.95	452.86	0.45	−0.65	15	>.001	−0.01
Pair 3 vs. Pair 4	13.78	0.82	452.86	0.45	−213.98	15	<.001	−3.38
Bin 5								
Pair 1 vs. Pair 2	994.15	0.21	554.25	3.72	63.01	15	<.001	1.15
Pair 1 vs. Pair 3	994.15	0.21	−4.88	0.84	457.80	15	<.001	8.36
Pair 1 vs. Pair 4	994.15	0.21	497.75	0.57	369.48	15	<.001	6.74
Pair 2 vs. Pair 3	554.25	3.72	−4.88	0.84	78.93	15	<.001	1.44
Pair 2 vs. Pair 4	554.25	3.72	497.75	0.57	14.60	15	<.001	0.27
Pair 3 vs. Pair 4	−4.88	0.84	497.75	0.57	−208.51	15	<.001	−3.81

Note: *M* = mean; CSE = clustered standard error; df = degrees of freedom.

APPENDIX G

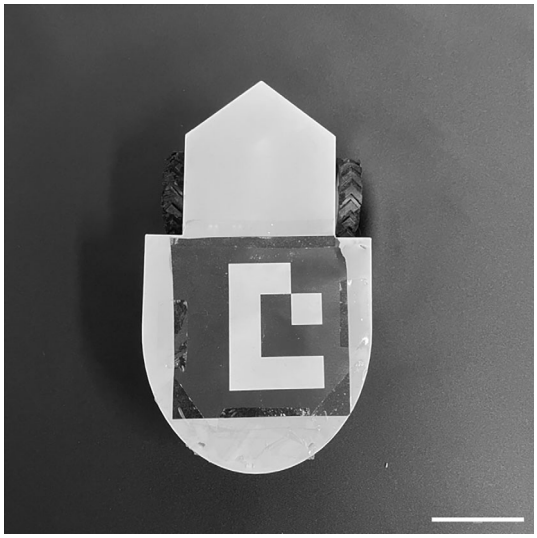

FIGURE G1 Photograph of the robot. The front wheels extend beyond the 3D-printed plastic shell (white), which is labeled with an ArUco marker to facilitate the robot's position estimation. Scale bar: 5 cm.

APPENDIX H

TABLE H1 Comparison of different pairs for rat reward, rat punishment, rat net reward, and robot reward.

					First pair		Second pair		
Comparison	**Test**	**df**	***F***	***p*** **(η^2)**	**M**	***SEM***	***M***	***SEM***	**Post hoc *p***
Rat reward	One-way ANOVA	7, 104	48.12	<.05 (0.76)					
Pair 1 vs. Pair 6	Tukey's post hoc test				8.29	0.81	0.29	0.13	<.05
Pair 1 vs. Pair 7	Tukey's post hoc test				8.29	0.81	0.14	0.10	<.05
Pair 2 vs. Pair 6	Tukey's post hoc test				8.57	0.51	0.29	0.13	<.05
Pair 2 vs. Pair 7	Tukey's post hoc test				8.57	0.51	0.14	0.10	<.05
Pair 3 vs. Pair 6	Tukey's post hoc test				9.14	0.48	0.29	0.13	<.05
Pair 3 vs. Pair 7	Tukey's post hoc test				9.14	0.48	0.14	0.10	<.05
Pair 4 vs. Pair 6	Tukey's post hoc test				7.64	0.45	0.29	0.13	<.05
Pair 4 vs. Pair 7	Tukey's post hoc test				7.64	0.45	0.14	0.10	<.05
Pair 5 vs. Pair 6	Tukey's post hoc test				8.29	0.62	0.29	0.13	<.05
Pair 5 vs. Pair 7	Tukey's post hoc test				8.29	0.62	0.14	0.10	<.05
Pair 6 vs. Pair 8	Tukey's post hoc test				0.29	0.13	8.50	0.81	<.05
Pair 7 vs. Pair 8	Tukey's post hoc test				0.14	0.10	8.50	0.81	<.05
Rat punishment	One-way ANOVA	7, 104	9.30	<.05 (0.38)					
Pair 1 vs. Pair 6	Tukey's post hoc test				3.36	0.58	0.71	0.27	<.05
Pair 1 vs. Pair 7	Tukey's post hoc test				3.36	0.58	0.43	0.14	<.05
Pair 2 vs. Pair 6	Tukey's post hoc test				3.00	0.35	0.71	0.27	<.05
Pair 2 vs. Pair 7	Tukey's post hoc test				3.00	0.35	0.43	0.14	<.05

TABLE H1 (Continued)

					First pair		Second pair		
Comparison	**Test**	**df**	***F***	***p*** **(η^2)**	**M**	***SEM***	***M***	***SEM***	**Post hoc *p***
Pair 3 vs. Pair 6	Tukey's post hoc test				2.79	0.49	0.71	0.27	<.05
Pair 3 vs. Pair 7	Tukey's post hoc test				2.79	0.49	0.43	0.14	<.05
Pair 4 vs. Pair 6	Tukey's post hoc test				4.07	0.52	0.71	0.27	<.05
Pair 4 vs. Pair 7	Tukey's post hoc test				4.07	0.52	0.43	0.14	<.05
Pair 5 vs. Pair 6	Tukey's post hoc test				3.93	0.63	0.71	0.27	<.05
Pair 5 vs. Pair 7	Tukey's post hoc test				3.93	0.63	0.43	0.14	<.05
Pair 6 vs. Pair 8	Tukey's post hoc test				0.71	0.27	3.64	0.50	<.05
Pair 7 vs. Pair 8	Tukey's post hoc test				0.43	0.14	3.64	0.50	<.05
Rat net reward	One-way ANOVA	7, 104	8.15	<.05 (0.35)					
Pair 1 vs. Pair 6	Tukey's post hoc test				4.93	1.30	−0.43	0.25	<.05
Pair 1 vs. Pair 7	Tukey's post hoc test				4.93	1.30	−0.29	0.13	<.05
Pair 2 vs. Pair 6	Tukey's post hoc test				5.57	0.80	−0.43	0.25	<.05
Pair 2 vs. Pair 7	Tukey's post hoc test				5.57	0.80	−0.29	0.13	<.05
Pair 3 vs. Pair 6	Tukey's post hoc test				6.36	0.86	−0.43	0.25	<.05
Pair 3 vs. Pair 7	Tukey's post hoc test				6.36	0.86	−0.29	0.13	<.05
Pair 4 vs. Pair 6	Tukey's post hoc test				3.57	0.88	−0.43	0.25	<.05
Pair 5 vs. Pair 6	Tukey's post hoc test				4.36	1.11	−0.43	0.25	<.05
Pair 5 vs. Pair 7	Tukey's post hoc test				4.36	1.11	−0.29	0.13	<.05
Pair 6 vs. Pair 8	Tukey's post hoc test				−0.43	0.25	4.86	1.18	<.05
Pair 7 vs. Pair 8	Tukey's post hoc test				−0.29	0.13	4.86	1.18	<.05
Robot reward	One-way ANOVA	7, 104	122.4	<.05 (0.89)					
Pair 1 vs. Pair 6	Tukey's post hoc test				11.64	0.53	1.00	0.33	<.05
Pair 1 vs. Pair 7	Tukey's post hoc test				11.64	0.53	0.57	0.20	<.05
Pair 2 vs. Pair 6	Tukey's post hoc test				11.57	0.34	1.00	0.33	<.05
Pair 2 vs. Pair 7	Tukey's post hoc test				11.57	0.34	0.57	0.20	<.05
Pair 3 vs. Pair 6	Tukey's post hoc test				11.93	0.45	1.00	0.33	<.05
Pair 3 vs. Pair 7	Tukey's post hoc test				11.93	0.45	0.57	0.20	<.05
Pair 4 vs. Pair 6	Tukey's post hoc test				11.71	0.42	1.00	0.33	<.05
Pair 4 vs. Pair 7	Tukey's post hoc test				11.71	0.42	0.57	0.20	<.05
Pair 5 vs. Pair 6	Tukey's post hoc test				12.21	0.59	1.00	0.33	<.05
Pair 5 vs. Pair 7	Tukey's post hoc test				12.21	0.59	0.57	0.20	<.05
Pair 6 vs. Pair 8	Tukey's post hoc test				1.00	0.33	12.21	0.66	<.05
Pair 7 vs. Pair 8	Tukey's post hoc test				0.57	0.20	12.21	0.66	<.05

Note: df = degrees of freedom; η^2 = effect size; *M* = mean; *SEM* = standard error of the mean.