Learning the rules of a game: neural conditioning in human-robot interaction with delayed rewards

Published in the *Proceedings of the Third Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics - Osaka, Japan - August 2013*

Andrea Soltoggio, Felix Reinhart, Andre Lemme and Jochen Steil
Research Institute for Cognition and Robotics (CoR-Lab) and Faculty of Technology
Bielefeld University, Germany.
Email: {asoltogg, freinhar, alemme, jsteil}@cor-lab.uni-bielefeld.de

***Abstract*—Learning in human-robot interaction, as well as in human-to-human situations, is characterised by noisy stimuli, variable timing of stimuli and actions, and delayed rewards. A recent model of neural learning, based on modulated plasticity, suggested the use of rare correlations and eligibility traces to model conditioning in real-world situations with uncertain timing. The current study tests neural learning with rare correlations in a human-robot realistic teaching scenario. The humanoid robot iCub learns the rules of the game rock-paper-scissors while playing with a human tutor. The feedback of the tutor is often delayed, missing, or at times even incorrect. Nevertheless, the neural system learns with great robustness and similar performance both in simulation and in robotic experiments. The results demonstrate the efficacy of the plasticity rule based on rare correlations in implementing robotic neural conditioning.**

I. Introduction

The interactive nature of learning in humans is suggested to be crucial from the very early stages of development [1], [2]. The complexity and rich dynamics of interactive learning processes constitute a challenge to the understanding of the underlying mechanisms. When a tutor gives feedback, e.g. an encouragement or a disincentive, it may not be entirely clear to what precise actions the feedback refers to. Time-delays and disturbances create a complex input-output pattern that makes it difficult to extract correct cause-effect relationships. Modelling such types of learning requires the acknowledgement of the asynchrony of input and output flows with imprecise timing and unreliable signals and actions. A focus in cognitive developmental robotics [3] is to model interactive feedback during learning.

Consider for instance a child engaging in playing the game rock-paper-scissors. Some of the required skills and rules, e.g. what are the possible moves and the timing of the moves, can be arguably acquired by observational learning [4]. By observing and imitating, a child can learn to perform the correct actions and play according to an agreed protocol. Understanding the rules, i.e. who wins and who loses, goes beyond simple imitation. Such higher level rules are generally conveyed verbally and explicitly from the expert to the novice. But the saying that *practice makes perfect* suggests that associations and mastering of the rules emerge only with performing, and possibly with errors.

In the scenario considered in this study, the success of a particular action is evaluated considering the feedback that follows. In the case of the game rock-paper-scissor, learning the rules may imply that a novice attempts to guess who wins and who loses, and adjusts that knowledge according to the feedback provided by an expert. Failure to classify correctly a move leads to negative feedback from the tutor, while success leads to positive feedback. Such type of learning can be classified in the overall category of operant conditioning [5], [6].

Instances in which a novice improves performance by adjusting his or her actions according to feedback are numerous. Social interaction [2], [7], motor learning [8], and other intelligent behaviours are examples off areas in which skills are refined by means of a complex flow of forward and feedback information. Interestingly, such feedback is all but precise. It is characterised by delays and uncertainties generated by disturbing stimuli, unreliable signals, or unfocused attention. In general, rewards and punishments are delayed and refer to actions that were performed previously in time. The problem of associating current rewards with previous actions was named *distal reward problem* [9], or temporal credit assignment [10] and it is ubiquitous in humans [11] and machine learning [12], [13]. In the latter, algorithms were developed to perform statistical analysis of events to find temporal associations [14]. In neural learning, however, an open question is how does learning occur if the neural activity generated by stimuli and actions is no longer present seconds later when a reward is perceived?

Relatively few studies focus on the neural mechanisms that bridge the temporal gap between sequences of cues, actions and rewards [15], [16], [17]. In [15], the precise spike-timing of neurons was indicated as the essential feature to perform classical and operant conditioning with modulated spike-timing-dependent plasticity. This position was challenged in a recent study [17] in which the *rarity* of both correlating neural activity and eligibility traces was identified as the main feature that allowed for the solution of the distal reward problem even without spiking neurons. The rarity of correlations was shown in simulation to be responsible for selecting rare neural events. Such events are then propagated further in time and enable to reconstruct which actions lead to rewards.

Neural operant conditioning is reproduced in the current study by means of the plasticity rule *rarely correlating Hebbian plasticity* (RCHP) [17], [18]. The rule prescribes the use of rare neural correlations in combination with modulated Hebbian plasticity to find correct associations between past actions and present rewards. The experiment in a robotic scenario validates the principle of rare correlations in implementing effective dynamics to solve the distal reward problem in realistic scenarios.

The game rock-paper-scissor is used in the current study as an exemplary instance in which the delayed feedback from the tutor is interpreted by a robot to learn the rules of the game. The interesting features of feedback interactive learning make robotic models an appealing platform for testing hypotheses on neural learning. The modelling of neural operant conditioning in real world conditions with delayed human-driven feedback is the focus of this study. The high level rules of the game, i.e. which combinations of moves lead to which results, are conveyed by the tutor to the robot by means of delayed and uncertain feedback. The robot learns by trial and error, uses the imprecise feedback from the tutor and demonstrates the improving skill while performing the game.

It is important to note that the speed of execution is not a focus, as well as the acquisition of winning strategies that is outside the scope and intent of this study. The robot and the tutor play the game at a moderate speed. The aim of the tutor is not that of winning, but rather that of teaching the iCub the rules of the game by providing positive and negative feedback. The robot was pre-programmed to execute the actions of the game in a similar fashion to a human player. This skill could be acquired by imitation learning [19], but was instead preprogrammed in the current setting to focus on the feedback reward learning. The complete robotic system is structured with a relatively complex feedback control architecture. The study, however, focuses on the demonstration of neural learning and plasticity to solve the distal reward problem, rather than on the building of complete control architectures [20]. The results demonstrate the neural learning of the rules of the game by trial and error, and show an effective model and implementation of operant conditioning.

The robotic setup and game, the learning network and plasticity rules are described in the next section. The results, showing the learning dynamics of the neural weights, are presented subsequently in section III. A discussion and conclusion follows in section IV. The study is supplemented with the Matlab scripts to run the experiments in simulations and additional visual material at the author's associated website (http://andrea.soltoggio.net/rps).

II. The Robotic Platform and the Learning Network

A robot is endowed with a learning neural model and is placed in an interactive scenario with the intent of learning the rules of the game rock-paper-scissors. This section describes the features of the robot, the inputs and outputs, and the details of the neural model.

A. The iCub plays rock-paper-scissors

The iCub is a child-sized humanoid robot of 90 cm of height, weighing 23 kg and comprising 53 degrees of freedom [21]. Its size and aspect make the iCub particularly suitable and conducive to interactive learning with humans. The eye cameras and tactile sensors can be used to receive a variety of stimuli and feedback from the interaction partners and from the environment. Head movements, gazing, speech and facial expressions endow the iCub with a particularly rich set of communication modalities.

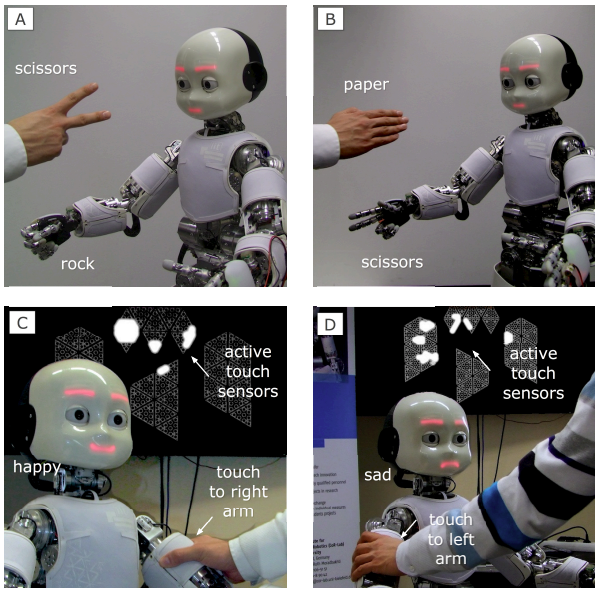

Fig. 1. Various phases of the interactive process of the iCub with a person. (A) The iCub is observing the move of the tutor after playing rock (the tutor played scissors). (B) The iCub observes the tutor's move, paper. (C) The tutor gives a reward by touching the iCub's left arm after the iCub has correctly classified the outcome of one trial. (D) The tutor gives a negative feedback by touching the iCub's right arm after it has incorrectly classified the outcome of one trial. In the background of photos C and D a monitor displays the sensing areas on the arms: the currently active touch sensors appear as lighted patches.

The movements of the robot were implemented to result in a natural-looking sequence while interacting with people. When a person approaches the robot, the iCub initiates the game by saying *Let's play*, which is immediately followed by the enunciation of *One, two, three, go*. At the word *one*, the iCub raises its right arm in preparation for the move. At the word *go* the iCub lowers its arm and adopts one of the three possible hand configurations. The iCub expects the human partner to perform the move simultaneously. At that moment, the iCub detects which move was performed by the opponent (Figs. 1A and 1B). Considering the tutor's move, and the iCub's own move, the robot attempts to guess the outcome of the trial. The three possible outcomes are enunciated as *I win*, *you win*, or *no one wins*. Such enunciations represent actions that are evaluated by the human tutor.

Initially the iCub does not know which combinations of moves is a win, lose, or even. The aim of the experiment is to investigate the capability of performing operant conditioning based on the tutor's feedback. Such a feedback is conveyed by the tutor by pressing the left or the right arm of the robot according to whether the robot provides correct or incorrect answers (Figs. 1C and 1D). The complete robotic and software architecture that performs the task is illustrated in Fig. 2. The learning process is implemented by means of a plastic network that solves the distal reward problem as described in the next section.

B. Rarely Correlating Hebbian Plasticity

The Rarely Correlating Hebbian Plasticity (RCHP)[17], [18] is a type of Hebbian plasticity that extracts only rare correlations from the neural activity. In combination with

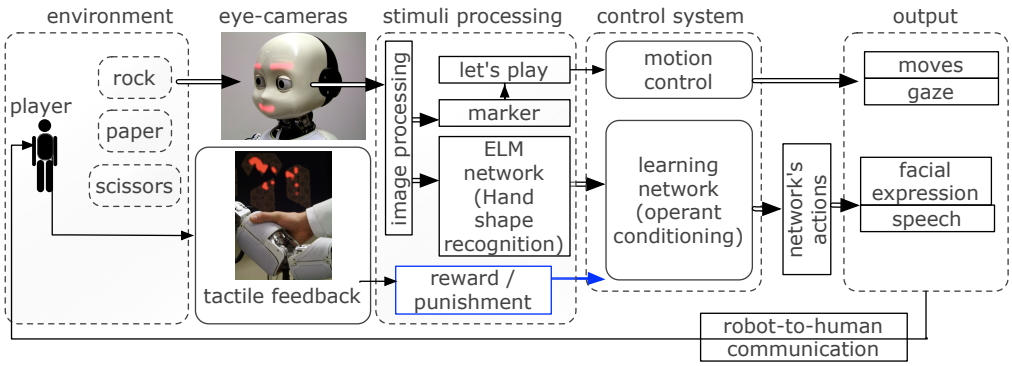

Fig. 2. Representation of the closed loop robotic system with the various components. The robot receives stimuli from the environment, performs actions and interacts with the human tutor. The inputs are in the form of visual images and tactile information. Visual stimuli are processed and analysed to determine which move is performed by the human player. The tactile information is read to received feedback. A pre-programmed motion-control block is responsible for performing the arm movements and hand configurations. The learning network that performs operant conditioning drives the robot to enunciate the outcome of each trial. Modelling the effect of tactile information as unconditioned stimulus, the robot displays positive and negative emotions through face expressions when receiving positive or negative feedback.

neuromodulation [22], it drives plasticity to reinforce reward-related stimuli and actions. The RCHP rule is given as

$$\mathrm{RCHP}_{ji}(t) = \begin{cases} +\alpha & \text{if } v_j(t-t_{pt}) \cdot v_i(t) > \theta_{hi} \\ -\beta & \text{if } v_j(t-t_{pt}) \cdot v_i(t) < \theta_{lo} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where j and i are a presynaptic and a postsynaptic neuron, α and β two positive learning rates (in this study set to 0.1) for correlating and decorrelating synapses respectively, $v(t)$ is the neural output, t_{pt} is the propagation time of the signal from the presynaptic to the postsynaptic neuron, and θ_{hi} and θ_{lo} are the thresholds that detect highly correlating and highly decorrelating activities.

The thresholds θ_{hi} and θ_{lo} are estimated online to target an average rate μ of approximately 0.5%/s of rare correlations [17]. θ_{hi} and θ_{lo} are assigned initially arbitrary values of 0.1 and -0.1 respectively. A first-in first-out queue $cq(t)$ holds the number of correlations between neurons registered at each step during the recent past (in this implementation for the last 10 s). If the number of measured correlations during the last 10 s is higher than 5 times the target μ, i.e. higher than 2.5%, θ_{hi} is increased by a small step $\eta = 0.002/s$. If the correlations are too few, i.e. less than $\frac{1}{5}\mu$ (0.1%), the threshold is decreased of the same small step. The same procedure is applied to estimate θ_{lo}. More details of the RCHP can be found in [17], [18].

Rare correlations create synapse-specific eligibility traces that work in combination with neuromodulation to solve the distal reward problem as explained in the next section.

C. The neural model for operant learning

The RCHP rule modifies synapse-specific eligibility traces c_{ji} between a presynaptic neuron j and a postsynaptic neuron i. Eligibility traces are slow-decaying values that preserve a memory of recent events. A modulatory signal m, which is governed by a fast decay and by the exogenous input reward $r(t)$, converts eligibility traces to weight changes. The changes of the eligibility traces c_{ij}, weights w_{ij} and modulation m are governed by

$$\dot{c}_{ji} = -c_{ji}/\tau_c + \mathrm{RCHP}_{ji}(t) \quad (2)$$

$$\dot{w}_{ji}(t) = m(t) \cdot c_{ji}(t) \quad (3)$$

$$\dot{m}(t) = -m(t)/\tau_m + \lambda \cdot r(t) + b \quad . \quad (4)$$

where a reward episode at time t sets $r(t) = 1$, which increases the value of $m(t)$ proportionally to a constant λ. A baseline modulation b can be set to a small value and has the function of maintaining a small level of plasticity. The modulatory signal decays relatively quickly with a time constant $\tau_m = 1$ s, while the time constant of eligibility traces τ_c is 4 s. The neural state u_i and output v_i of a neuron i are computed with a rate-based model expressed by

$$u_i(t) = \sum_j (w_{ji} \cdot v_j(t) \cdot \kappa_j) + S_i \quad (5)$$

$$v_i(t+\Delta t) = \begin{cases} \tanh\big(\gamma \cdot u_i(t)\big) + \xi_i(t) & \text{if } u_i \geq 0 \\ \xi_i(t) & \text{if } u_i < 0 \end{cases} \quad (6)$$

where w_{ji} is the connection weight from a presynaptic neuron j to a postsynaptic neuron i; κ_j is $+1$ and -5 for excitatory and inhibitory neurons respectively to reflect the stronger effect of less numerous inhibitory neurons; S_i represents the increase of u by external stimuli; γ is a gain parameter; $\xi_i(t)$ is a uniform noise source drawn in the interval [-0.1,0.1]. The sampling time is set to 200 ms, which is also assumed to be the propagation time t_{pt} (Eq. 1) of signals among neurons.

The network has 800 neurons: 620 are excitatory and 180 inhibitory. Their activity and outputs are governed by Eqs. 5 and 6. Each neuron is connected to another neuron with probability 0.1. All excitatory neurons have plastic afferent connections that vary according to Eq. 3 and are constrained by saturation in the interval [0,1]. Inhibitory neurons have fixed afferent connections. The network has therefore a random connectivity and random initial weights. Input signals are conveyed to the network by means of groups of randomly selected neurons, each group containing 50 neurons. When one input is active, the neural states u_i of the neurons in the corresponding group increase by 10, i.e. $S_i = 10$ in Eq. 5. Other groups of randomly selected neurons (50 each group) are elected as output groups and their activities decide actions.

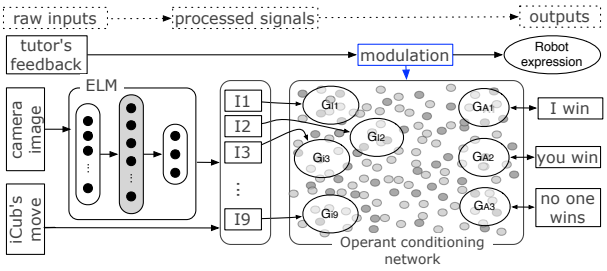

Fig. 3. Graphical representation of the learning networks (expansion of central part of Fig. 2). The ELM network is responsible for the hand gesture recognition and is trained before the game starts. Two online learning network is composed of 800 neurons. The binary stimuli $I1..I9$ indicate one of the nine possible configurations of the game as illustrated in Table I. The stimuli are delivered to their respective groups of randomly chosen neurons $G_{I1}..G_{I9}$. The groups $G_{A1..A3}$ determine the enunciation of one particular sentence. The haptic sensor delivers a reward that represents the unconditioned stimulus (US).

iCub/tutor	rock	paper	scissors
rock	no one wins (I1)	you win (I2)	I win (I3)
paper	I win (I4)	no one wins (I5)	you win (I6)
scissors	you win (I7)	I win (I8)	no one wins (I9)

TABLE I. OUTCOMES OF ALL POSSIBLE MOVES.

D. Stimuli and actions

The robot perceives the moves of the opponent interpreting the camera image from the cameras in the eyes. A skin-colour filter with a subsequent neural classifier was implemented with a variant of the multi-layer perceptron called extreme learning machine (ELM) [23]. In a preliminary learning phase, the iCub was shown examples of rock, paper, and scissors to train the weights of the ELM with supervised learning.

The robot was programmed to execute the moves rock, paper, or scissor, and expected to detect the opponents move at the end of the sequence of movements. At each trial, the iCub played a random move and detected which move the human partner performed (rock, paper or scissor). This information was fed to the learning network as input.

If the iCub could successfully identify the move of the opponent, the corresponding input to the operant conditioning network was activated (Fig. 3). The inputs are binary values that represent which particular combination (out of 9) is detected (Table I). At the moment of the activation of one input, the output groups are monitored to perform an action according to which group has the highest activity. The output group that performs the action becomes then slightly activated by means of an action-to-network feedback which increases the activation u of the output neurons by 1. This feedback has the purpose of expressing, in terms of neural activation, which action was performed.

III. EXPERIMENTAL RESULTS

The iCub was programmed to engage immediately in the game as soon as it saw someone standing in front. One trial occurred approximately every 6 to 12 s. Throughout the trials, the network adjusted the weights driven by the feedback and improved the performance over time. There are 27 pathways that connect nine input neuron groups to three output neuron groups. Of those, nine pathways, one for each input group

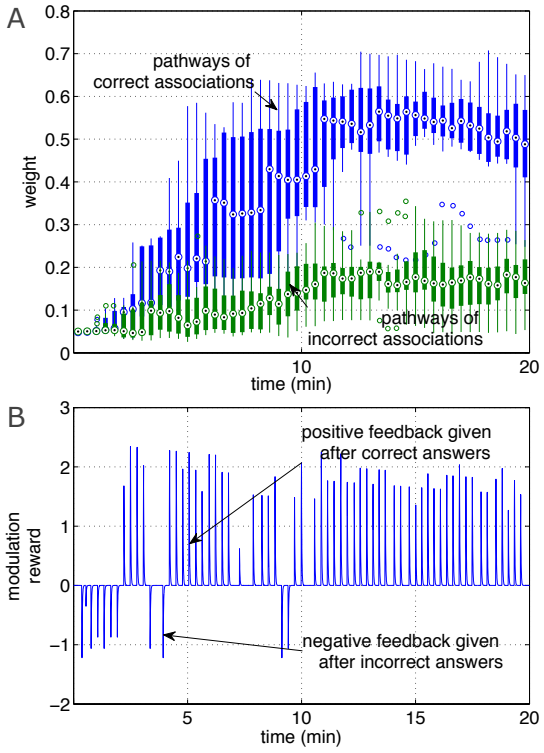

Fig. 4. Results of the robotic experiment. (A) Box plots showing the weights belonging to pathways that identify correct association (blue lines) and incorrect associations (green lines). Pathways leading to correct associations are observed to grow (blue lines) whereas pathways leading to incorrect associations remain at low values (green lines). (B) Reward (expressed as the modulatory signal m(t) given by the human tutor to the robot.

(see Table I), grew larger because represented the correct relationships between hand configurations and the outcome of the trial. The other 18 pathways represented incorrect associations and were not reinforced. Fig. 4A shows the statistics of the weight values grouped in the two categories, correct and incorrect associations. The statistics are illustrated by box plots indicating the median (central point), 25th and 75th percentiles (thick lines), most extreme data points (thin lines), and outliers (circles) [24]. The plot shows that, after 20 minutes, the pathways that lead to correct associations (blue lines) are significantly stronger than the pathways that lead to wrong associations (green lines). Fig. 4B shows the feedback received by the robot during one run. Initially, the tutor gave negative feedback a number of times because the robot made mistakes. After some time, the robot started to answer progressively more correctly, thereby proving that the rules of the game were successfully learnt. Fig. 4B additionally shows that the feedback provided by the tutor varies in absolute value. The reason is that the feedback signal is perceived as tactile information: variations in the pressure and length of time resulted also in variations of the total amount of signal perceived.

The associations between the moves of the game and correct answers is learnt despite considerable noise and disturbances in the overall system. Fig. 5 shows the neural activity of the three output groups (averaged over all the neurons in each group and represented with different colours). The

data discloses two important characteristics of the system. 1) The neural activity that triggered a response, visible in the peaks in Fig. 5, is not present anymore at the moment of the feedback. 2) The peaks that triggered the action, although clearly visible in the graph, are preceded and followed by a constant and non-negligible level of neural activity. These two features intuitively suggest that learning which pathways lead to a reward is not a trivial problem. The RCHP rule, by selecting highly correlating activity and generating slow-decaying traces, represents a plasticity mechanism particularly suitable to solve the current problem.

The RCHP rule is capable of learning also when intervening and disturbing stimuli are perceived before the reward, as demonstrated in [17], [18], if learning rates are sufficiently low. In the particular setting of this study, no disturbing actions or stimuli occurred before reward delivery. However, the considerable level of spontaneous and noisy activity that intervenes between actions and rewards implies that a reward does not follow immediately the action. In effect, the neural activity encodes input and output signals in a sub-symbolic, asynchronous and noisy pattern. Thus, there is no relation in the network between actions and following rewards except for a relative proximity in time and the eligibility traces created by the RCHP. Note that the traces are not associated to stimuli or actions, but are a property of synapses, which in turn are randomly placed in the network. The problem of mapping 9 inputs to 3 outputs is per se trivial. The difficulty is created by the need for a relatively fast reaction of the output, i.e. the robot's answers, and the delayed reward after several seconds, which may or may not occur and is of variable intensity. To the best of the authors' knowledge, such a problem has not been previously solved by randomly connected networks with high levels of spontaneous activity, and with a sampling time that is faster and unrelated to the timing of stimuli and actions, and to the delay of the feedback.

The rarity of correlations of the RCHP was set to 0.5% of correlations/s, as prescribed in [17], [18], in combination with traces with 1-3 s time constants that can account for feedback with delays within 10 s. Learning can be maintained with longer delays if the rarity of correlations is further decreased as shown in the cited literature. The possibility of setting the RCHP rule to account for longer delays or more disturbing stimuli is particularly suited to robotic scenarios as the one presented here.

The experiment conducted with the robot was also performed with simulated inputs and outputs without a robot. The moves of both the tutor and of the robot were simulated by drawing random combinations. An automated agent gave feedback to the network with a variable delay in the interval $[0, 5]$ s. Fig. 6A shows the box plots of the pathways for the correct and incorrect associations throughout 1 h of simulated time. Ten independent runs were performed to observe reliable statistical data. In the simulation, as opposed to the real robot experiment (Fig. 4A), the separation between pathways is more evident, i.e. the learning appears to identify more clearly the correct pathways. It must be noted that the robot experiment run for 20 m, while the simulations were run for the longer period of 1 h. The feedback provided by the code in simulation, although had a variable delay, was more reliable than the feedback provided by the human tutor. In some occasions, the

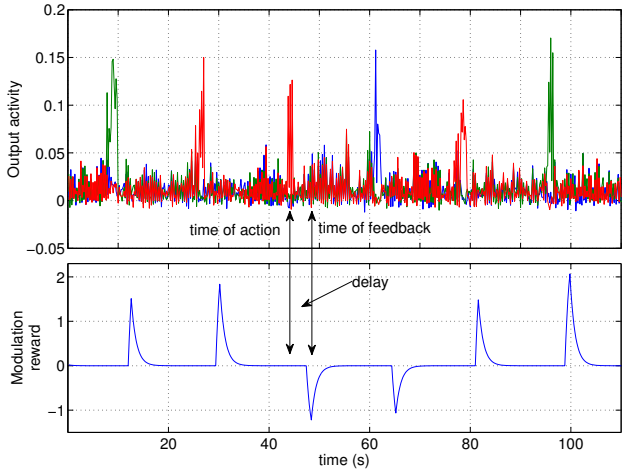

Fig. 5. Close-up on the neural activity of the output groups G_{A1}, G_{A2} and G_{A3}. The average activity of the 50 neurons of each output group (Fig. 3), computed as $1/50 \sum v_i \forall i \in G_{Ax}$, is shown with three different colours in the top graph. The bottom graph shows the modulatory activity that reflects the feedback provided by the tutor. Due to noise and spontaneous activity, the output neurons are not silent during waiting time, i.e. when no action is performed. Both plots show that the feedback is delivered after the action is performed and when the activity of the network does not reflect anymore the action previously performed.

tutor gave incorrect feedback, or waited a long time to provide it. Additionally, the timing of signal perception and action execution is further disturbed by delays in the communication network (LAN) that connected the robot with the computers performing the computation. Package loss also determined in a few occasions an irregular or unpredictable behaviour of the robot. Technical issues in signal propagation, failure in detecting correct stimuli and the unreliability of human feedback are elements that support the value of successful learning in this scenario, and demonstrate the robustness and reliability of the learning network. Fig. 6B shows the feedback signal during one simulated run. The efficient performance of operant conditioning determined only few initial negative reward episodes. After the initial learning phase, all feedback was positive, indicating that the learning network had successfully acquired all the rules of the game. In the 10 runs, the network learnt the correct rules after an average of 15 errors, with a worse performance in one run with 22 errors and a best performance in one run with 8 errors. Note that the amount of reward decreased slowly throughout the simulation. This setting is useful to decrease progressively the amount of plasticity in the network as learning takes place. A decreasing reward is also biologically plausible because it represents a form of habituation.

IV. Conclusion

This study demonstrated the feasibility of neural operant learning in a human-robot interactive scenario. The rules of the game rock-paper-scissors were learnt throughout a process of trial and error. The learning process was guided by human feedback and was characterised by imprecise timing of stimuli, actions, and delayed rewards. Such conditions constitute a challenge for models of neural learning because correct associations among stimuli, actions, and rewards are difficult

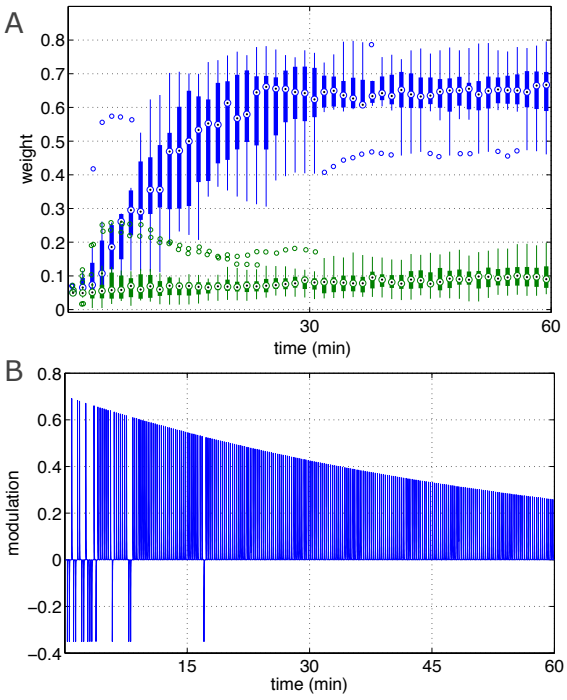

Fig. 6. Statistical analysis in simulation and example of feedback. (A) As in Fig. 4, the pathways leading to correct associations (blue box plots) and those leading to incorrect associations (green box plots) are grouped and analysed statistically. The values of the pathways from 10 independent runs (90 pathways for the correct associations, 180 for the incorrect associations) are shown over a 1 h simulated time. (B) Modulation received by the network in one particular simulation.

when such events are asynchronous, delayed or inconsistent. The analysis revealed that the neural activity at the moment of reward did not reflect the past actions, underlying the essential role of the rarely correlating Hebbian plasticity (RCHP). Human feedback was unreliable and characterised by variable intensity and delays. The plasticity model based on the RCHP created eligibility traces that allowed the network to overcome timing problems and derive correct associations. The successful learning was demonstrated both in simulation and with the real robot iCub. Using positive and negative feedbacks for the operant conditioning resulted in the convergence to stable correct behaviour in a low number of trials.

The learning scenario presented in this study is particular conducive to test hypotheses of learning in interaction. Learning that uses observation, imitation, and feedback is characterised by imprecise pairing of events. Human and animal intelligence appears to overcome discrepancies in timing and disturbances. The current study demonstrates the suitability of the RCHP rule to extend neural models with the capability of performing and predicting operant conditioning in realistic scenarios.

Acknowledgment

This work was supported by the European Communitys Seventh Framework Programme FP7/2007-2013, Challenge 2 Cognitive Systems, Interaction, Robotics under grant agreement No 248311 - AMARSi.

References

[1] B. Rogoff, C. Malkin, and K. Gilbride, "Interaction with babies as guidance in development," *New Directions for Child and Adolescent Development*, vol. 1984, pp. 31–34, 1984.

[2] M. Gauvain, *The social context of cognitive development*. Guilford Press, 2000.

[3] M. Asada, K. F. MacDormanb, H. Ishigurob, and Y. Kuniyoshic, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, pp. 185–193, 2001.

[4] A. Bandura, *Social Lerning Theroy*. New York: General Learning Press, 1977.

[5] E. L. Thorndike, *Animal Intelligence*. New York: Macmillan, 1911.

[6] B. F. Skinner, *Science and Human Behavior*. New York, MacMillan, 1953.

[7] L. Vygotsky, *Mind in Society*. Cambridge, MA: Harvard University Press, 1978.

[8] R. A. Schmidt, *Motor control and learning*. Champaign, III. Human Kinetics, 1982.

[9] C. L. Hull, *Principles of behavior*. New-Your: Appleton Century, 1943.

[10] R. S. Sutton, "Temporal credit assignment in reinforcement learning," Ph.D. dissertation, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, 1984.

[11] W. S. Millar and J. S. Watson, "The effect of delayed feedback on infant learning reexamined," *Child Development*, vol. 50, no. 3, pp. 747–751, 1979.

[12] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1998.

[13] E. Alpaydn, *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. MIT Press Cambridge, MA, USA, 2004.

[14] P. R. Cohen, C. Sutton, and B. Burns, "Learning effects of robot actions using temporal associations," in *Development and Learning, 2002. Proceedings. The 2nd International Conference on*. IEEE, 2002, pp. 96–101.

[15] E. M. Izhikevich, "Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling," *Cerebral Cortex*, vol. 17, pp. 2443–2452, 2007.

[16] M. Päpper, R. Kempter, and C. Leibold, "Synaptic tagging, evaluation of memories, and the distal reward problem." *Learning & Memory*, vol. 18, pp. 58–70, 2011.

[17] A. Soltoggio and J. J. Steil, "Solving the Distal Reward Problem with Rare Correlations," *Neural Computation*, vol. 25, no. 4, pp. 940–978, 2013.

[18] A. Soltoggio, A. Lemme, F. R. Reinhart, and J. J. Steil, "Rare neural correlations implement robotic conditioning with reward delays and disturbances," *Frontiers in Neurorobotics*, vol. 7, no. Research Topic: Value and Reward Based Learning in Neurobots, 2013.

[19] T. Chaminade, E. Oztop, G. Cheng, M. Kawato *et al.*, "From self-observation to imitation: Visuomotor association on a robotic hand," *Brain research bulletin*, vol. 75, no. 6, pp. 775–784, 2008.

[20] P. Andry, A. Blanchard, and P. Gaussier, "Using the rhythm of nonverbal human–robot interaction as a signal for learning," *Autonomous Mental Development, IEEE Transactions on*, vol. 3, no. 1, pp. 30–42, 2011.

[21] N. Tsakarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, F. Becchi, L. Righetti, J. Santos-Victor, A. Ijspeert, M. Carrozza, and D. Caldwell, "iCub - The Design and Realization of an Open Humanoid Platform for Cognitive and Neuroscience Research," *Journal of Advanced Robotics, Special Issue on Robotic platforms for Research in Neuroscience*, vol. 21, no. 10, pp. 1151–1175, 2007.

[22] A. Soltoggio and K. O. Stanley, "From Modulated Hebbian Plasticity to Simple Behavior Learning through Noise and Weight Saturation," *Neural Networks*, vol. 34, pp. 28–41, October 2012.

[23] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme Learning Machine: A new learning scheme of feedforward neural networks," in *IEEE Intern. Joint Conf. on Neural Networks*, 2004, pp. 985–990.

[24] R. McGill, J. W. Turkey, and W. A. Larsen, "Variations of box plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, February 1978.