Stabilising Hebbian learning with a third factor in a food retrieval task

Adedoyin Maria Thompson[1], Bernd Porr[1], and Florentin Wörgötter[2]

[1] Department of Electronics & Electrical Engineering, University of Glasgow, Glasgow, G12 8LT, Scotland, United Kingdom {mariat,b.porr}@elec.gla.ac.uk
[2] Bernstein Center of Computational Neuroscience, University Göttingen, Germany, worgott@chaos.gwdg.de

Abstract. When neurons fire together they wire together. This is Donald Hebb's famous postulate. However, Hebbian learning is inherently unstable because synaptic weights will self amplify themselves: the more a synapse is able to drive a postsynaptic cell the more the synaptic weight will grow. We present a new biologically realistic way how to stabilise synaptic weights by introducing a third factor which switches on or off learning so that self amplification is minimised. The third factor can be identified by the activity of dopaminergic neurons in VTA which fire when a reward has been encountered. This leads to a new interpretation of the dopamine signal which goes beyond the classical prediction error hypothesis. The model is tested by a real world task where a robot has to find "food disks" in an environment.

1 Introduction

Hebbian learning [1] is the most prominent paradigm in correlation based learning: If pre- and postsynaptic activity coincides the weight of the synapse is strengthened. However, Hebbian learning is inherently unstable because of its *autocorrelation* term: Briefly, a changing weight will alter the output which will lead to further weight change, and so on. In this study we present a novel learning rule which is an extension of our differential Hebbian learning [2] rule ISO-learning [3] which minimises the destabilising autocorrelation term by switching learning on when the autocorrelation term is minimal. This switching is performed by a third factor which acts like a neuromodulator [4]. Consequently we call this learning rule ISO3 learning because it is ISO learning with a third factor. We will demonstrate the applicability of the rule with a robot that learns to retrieve "food disks".

2 Three factor learning

We are going to demonstrate in the open loop case how to minimise the destabilising autocorrelation term of Hebbian learning. Fig. 1A shows the basic components of the neural circuit. The learner consists of three inputs x_0, x_1 and r which

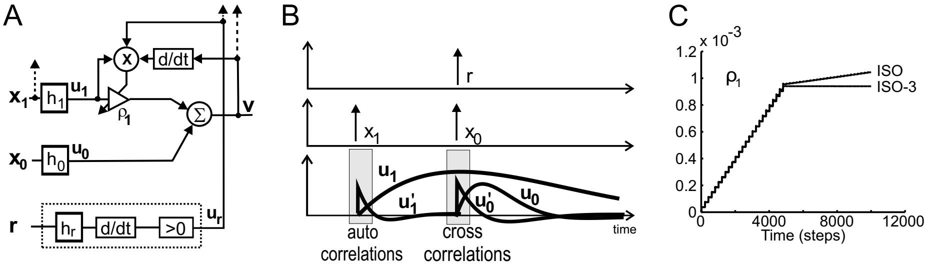

Fig. 1. A) General form of the neural circuit in a generic environment. The inputs x_0, x_1, r are filtered by standard resonators (h_0, h_1, h_r which have frequency f and quality Q as parameters) which smear out an input signal for about $1/f$ samples. u_0 and u_1 are summed at v with weights ρ_0 and ρ_1. The number of filters in the x_1 pathway can be extended to a filterbank with different resonators h_k and corresponding weights ρ_k which is indicated by the dotted lines. From the output of the filter h_r the derivative d/dt is taken and then rectified (> 0). The symbol $\otimes$ is a correlator and $\sum$ is a summation node. B) Signals u_0, u_1 and their derivatives which illustrate how learning works (see text for explanation). C) Comparing ISO and ISO3 learning rules. System parameters: $f_{h_0,h_1,h_r} = 0.1$ and damping parameter $Q = 0.51$ was used to filter inputs x_0, x_1 and relevance signal r. Learning rate was $\mu = 0.005$ for ISO learning rule and $\mu = 0.07$ for ISO3 rule. Time difference between x_1 and x_0 was $T = 10$ (x_1 always precedes x_0).

are filtered by low pass filters: $u_0 = x_0 * h_0$, $u_1 = x_1 * h_1$ and $u_r = \Theta((r * h_r)')$ where Θ is a threshold for > 0 as depicted in Fig. 1. The low pass filters smear out the input signals in time to generate appropriate motor responses. The circuit can easily be extended to a bank of filters with different resonators $h_j, j > 0$ and individual weights $\rho_j, j > 0$ to generate complex shaped responses [5]. The learning rule for the weight change $\frac{d}{dt}\rho_j$ is given as:

$$\rho_j' = \mu u_r u_j v', \; j > 0 \tag{1}$$

which is essentially ISO learning where we have added a third factor u_r.

To get a better understanding how the third factor u_r influences learning we split Eq. 1 into a superposition of a cross-correlation cc_j and an auto-correlation ac_j, multiplied by the third-factor u_r:

$$\rho_j' = \left(\underbrace{\rho_0 u_j u_0'}_{cc_j} + \underbrace{u_j \sum_{k=1}^{N} u_k' \rho_k}_{ac_j} \right) u_r \tag{2}$$

$$= (cc_j + ac_j) u_r \tag{3}$$

The cross-correlation cc_j drives learning by relating *different* inputs with each other so that, for example, in the case of simple conditioning the correlation of the conditioned stimulus (CS) and the unconditioned stimulus (US). Here the unconditioned input is x_0 which is smeared out in time by the filter h_0 and enters Eq. 2 in form of the signal u_0. The conditioned input x_1 enters Eq. 2 via a filterbank $h_j, j > 0$ and generates a number of different temporal traces u_j which are then correlated with u_0. Hence, the signals u_j are cross-correlated with the signal u_0. The autocorrelation term ac_j is the unwanted contribution to learning because it is correlating the conditioned responses (x_1) with themselves which lead to self-amplification of the corresponding weights.

To demonstrate how the third factor stabilises learning we generate artificially input signals x_0, x_1, r to our open loop circuit which are delta pulses (pulses that last for one unit step) that trigger damped filter responses (see Fig. 1B). It can be clearly seen that the autocorrelation ac and cross correlation terms cc happen at *different moments* in time. Consequently we can switch on learning when the autocorrelation is minimal and the cross correlation is maximal. This can be achieved by switching on the third factor u_r at the same time as the signal x_0 is triggered.

Fig. 1C shows the behaviour of ISO3 learning as compared to ISO-learning for a relatively high learning rate. To test the effect of the autocorrelation we switched off the signal x_0 after step 4000 which effectively removes the cross correlation. As shown in [3], at least for low learning rates in ISO-learning, the weights should stabilise after x_0 has been switched off. Instead, clearly one sees that ISO-learning contains an instability, which leads to an upward bend. This is different for ISO3 learning which does not contain this instability because learning is switched off when self amplifying autocorrelation terms would destabilise learning. ISO3 learning is also stable when there is a bank of filters in the x_1 pathway and/or when the filter functions are not orthogonal to each other because the autocorrelation is zero at the moment the third factor u_r is triggered.

In summary ISO3 learning uses the fact that auto- and cross correlation happen at different moments in time. Consequently we can stabilise differential Hebbian learning by switching learning on at the moment when the autocorrelation term is minimal.

3 Closed loop

The behavioural experiments of this section have two purposes: They will give the signals x_0, x_1 and r a behavioural meaning as well as demonstrate the superiority of ISO3 compared to ISO learning. We will present a task where a robot has to learn to retrieve "food disks" [6, 7]. This task will first be used for benchmarking and will then be demonstrated in a real robot. The robot has to find "food disks" from the distance. Initially the robot has only a pre-wired reflex which enables it to react to "food disks" at close range only. During learning this reflex reaction is correlated with distant stimuli which enable the robot to target "food disks" from the distance. In the simulation, we use sound and vision for distant and proximal stimuli which respectively replace the artificial input signals x_1, x_0 originally used in our open loop circuit. In the real robot experiment these two

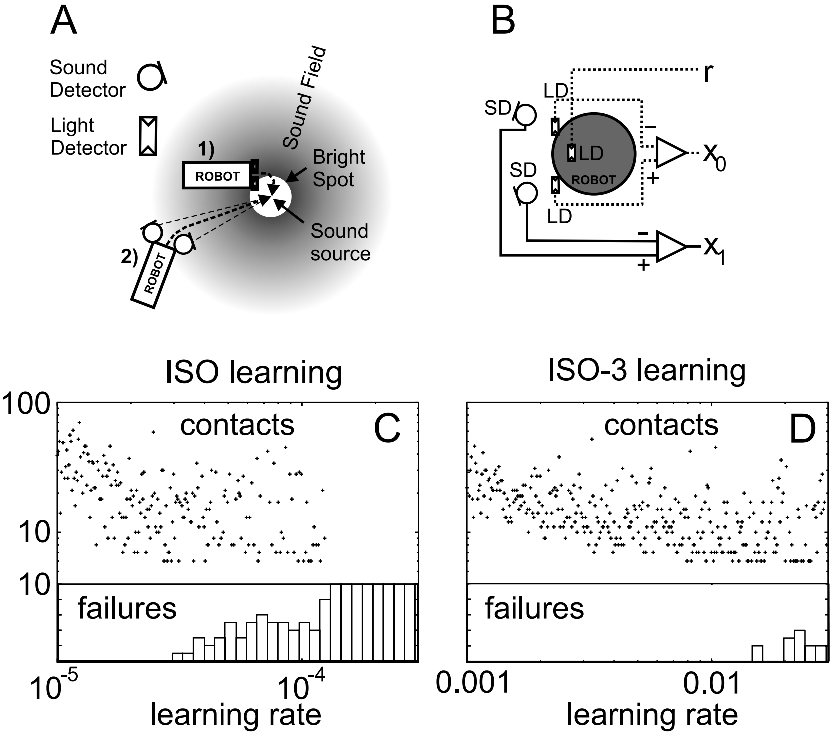

Fig. 2. The robot simulation. A) The robot has two pairs of sensors: It has two light sensors which detect the "food disk" only in their direct proximity. In addition it has two sound detectors which are able to "hear" the food source from a distance. B) The output v is the steering angle of the robot. The two light detectors (LD) establish the reflex reaction (x_0). The sound detectors (SD) establish the predictive loop (x_1). The weights $\rho_1 \ldots \rho_N$ are variable and are changed either by ISO or ISO3 learning. The signal r is generated by a third light sensor and is triggered as soon as the robot enters the "food disk". The robot has also a simple retraction mechanism when it collides with a wall ("retraction") which is not used for learning. The output v is the steering angle of the robot. Filters are set to $f_0 = 0.01$ for the reflex, $f_j = 0.1/j, j = 1 \ldots 5$ for the filter bank where $Q = 0.51$. Reflex gain was $\rho_0 = 0.005$. C) and D) plot the number of contacts for both learning rules needed for successful learning against the learning rate. In addition the number of failures against the learning rate are plotted.

signals x_1, x_0 will be generated from two different scanlines from a video camera attached to the robot.

3.1 Benchmark

Fig. 2A,B presents the task and circuit diagram where the simulated robot had to learn to retrieve "food disks". The reflex x_0 is established by two light detectors (LD) which draws the robot into the centre of the "food disks" (Fig. 2A1). Learning uses the sound detectors (SD, Fig. 2A2) which feed into x_1 to generate an anticipatory reaction towards the "food disk". The reflex reaction is established by the *difference* of two light dependent resistors which cause a steering reaction towards the white disk (Fig. 2B). Hence x_0 is equal to zero if both LDs are not stimulated or when they are stimulated *at the same time* which happens during a straight encounter with a disk. The latter situation occurs after successful learning. The reflex has a constant weight ρ_0 which always guarantees stable behaviour. The predictive signal x_1 is generated by using two signals coming from the sound detectors (SD). The signal is simply assumed to give the Euclidean distance from the sound source. The difference of the signals from the left and the right microphone is a measure of the azimuth of the sound source to the robot. Successful learning leads to a turning reaction which balances both sound signals and results ideally in a straight trajectory towards the target disk ending in a head-on contact.

We quantify successful and unsuccessful learning for increasing learning rates μ. The learning rates have been chosen in a way that in both cases the contacts for successful learning are the same to make the failures comparable. Learning was considered successful when we received a sequence of five contacts with the disk at a sub-threshold value of $|x_0| < 1.1$. We recorded the actual number of contacts until this criterion was reached. The simulations demonstrate clearly that ISO3 learning is much more stable than the Hebbian ISO learning. ISO3 learning can therefore operate at more than ten times higher learning rates than ISO learning.

3.2 Real robot

In this section we will demonstrate that ISO3-learning is also able to master the task with the "food-disk" in a physically embodied agent [8]. It will also be shown that ISO learning fails here completely because of its destabilising autocorrelation terms which drive the weights either very quickly to infinity or, alternatively, one has to run the robot for hours to see anticipatory behaviour which is impractical.

As before, the task of the robot is to target a white disk or "food disks" from a distance. As in the simulation the robot has a reflex reaction which pulls the robot into the white disk just at the moment the robot drives over the disk (Fig. 3). This reflex reaction is achieved by analysing the bottom scanline of a camera with a fisheye lens mounted on the robot. The predictive pathway is created in a similar way: A scanline which views the arena at a greater distance from the robot (hence "in its future") is fed into a bank of of ten filters. This enables the robot to learn to drive *towards* the "food disk" (Fig. 3).

The reflex behaviour of the robot before learning is shown in Fig. 4A, where the robot drives in a straight line and only makes a sharp bend when it encounters the "food disk" in very close proximity. i.e. when the "food disks" appears in the scanline that represents objects closest to the robot. Learning needs longer in these real robot runs than in the simulation. After about 5 minutes, the robot starts exhibiting a learned behaviour. Successful learning can be shown in Fig. 4B and C where the robot's turning reaction sets in from a distance of about 40cm. The robot has learned anticipatory behaviour.

The real robot is subject to complications which do not exist in the simulation. The inertia of the robot, imperfections of the motors and noise from the camera render learning more difficult than in the simulation. These elements contribute to the fluctuations in the weight change in Fig. 4D. The weights however remain stable. The two slightly large "jumps" (marked by circles) in the weight change between time steps 18000 . . . 20000 and 40000 . . . 45000 have been caused by typical problems which arise in real robots which have been mainly reflections on the floor and also the erroneous detection of the hand of the operator which caused weight changes. However, learning does not diverge and further learning makes the weights decrease again which points to the fact that the reflex reaction kicks in and corrects the slightly too strong steering reactions.

In order to fully appreciate the overall effects of the third factor, we have ran a real robot experiment implementing ISO learning without the third-factor by setting $u_r = 1$ all the time in Eq. 1. The learning rate has been reduced so

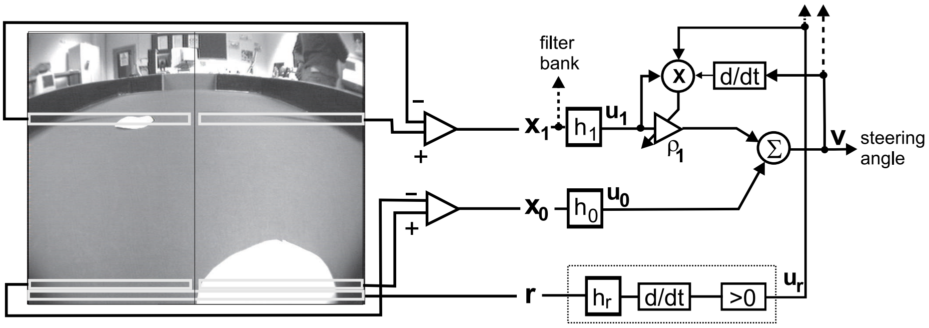

Fig. 3. The real robot's perspective showing two instances where the "food disks", represented by the white spheres lie in all scanlines at which the x_1, x_0 and the relevance input signals are established to produce the input signals for the learning circuit. The x_1 and x_0 signals are respectively triggered when the "food disk" appears in the upper scanline, where objects are further away from the robot's camera view and the lower scanline at bottom of the video image, where objects are closer to the robot's camera view. The relevance signal is obtained from the same scanline as the x_0 signal. When the "food disk" appears in either scanlines for the respective x_1, x_0 and relevance signal, a positive negative or zero value is generated depending on what side of the robots view the "food disks" lie. Parameters: frame rate was 25 frames/sec. The video image $f(x = [0 \ldots 95], y = [0 \ldots 64])$ was evaluated at $y = 53$ for the reflex x_0 and at $y = 24$ for the predictive signal x_1. Reflex and predictive signal were calculated as a thresholded (> 240) weighted sum: $x_{0,1} = \sum_{x=0}^{95} (x - 96/2)^2 \Theta(f(x, y))$. The reflex pathway was set to: $f_0 = 0.01, Q = 0.51$ with a reflex gain of $\rho_0 = 30$.The relevance filter was set to $f_r = 0.01, Q = 0.51$. The predictive filters were set to $f_1, k = 0.1/k, k \ldots 10, Q = 1$. The learning rate was $\mu = 0.0000035$.

that the weight development under ISO learning is comparable with ISO3 (see Fig. 4D). The weight change generated from this experiment is shown in Fig. 5. It can clearly be seen that ISO learning becomes unstable very quickly. Only after 2500 frames the weights diverge which leads to random behaviour of the robot so that the experiment was aborted.

In summary it can be concluded that ISO3 learning is much more stable than ISO learning: While ISO3 learning learns fast and remains stable, ISO learning diverges very quickly. This shows that the elimination of the autocorrelation term in ISO3 creates fast and reliable learning.

4 Discussion

In this work we have shown that a third factor is able to stabilise differential Hebbian learning by switching it on when its autocorrelation term is minimal.

Our ISO3-learning rule seems to have similarities with reinforcement learning (RL) which also employs a modulatory signal to select actions [9, 10]. However,

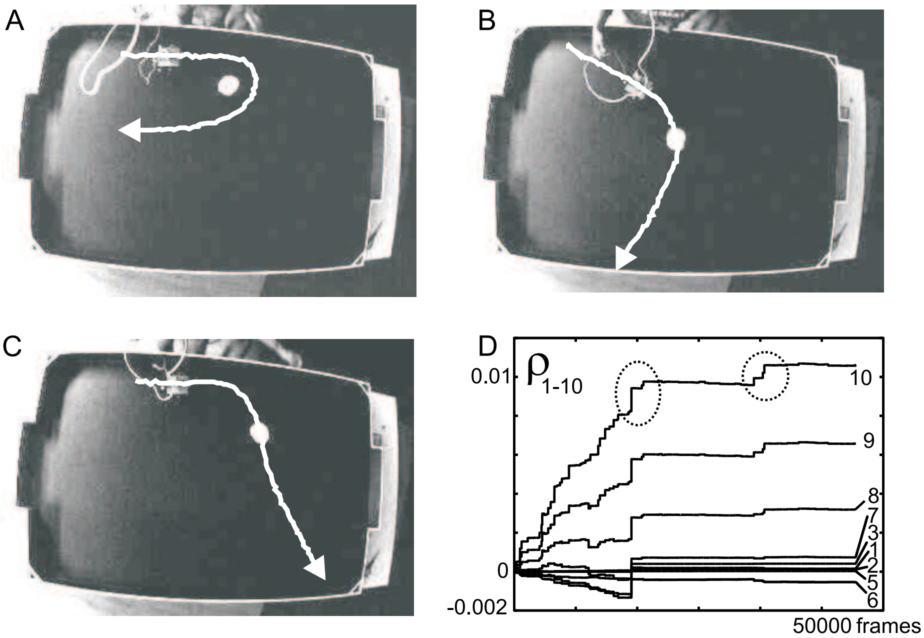

Fig. 4. Experiment with a real robot. A: start of the run at 00:12 mins, B: after 16:13 mins (92 contacts) weight change at a time step of approximately 24000, and C:after 24:10 mins (132 contacts) and the weight change at an approximate time step of 37000. The arrows at A and B show the trace of the robot while driving into "food disks" (white spheres). The weight development ($\rho_j, j = 1 \ldots 10$) is shown in D. The film can be viewed at `http://www.berndporr.me.uk/films`.

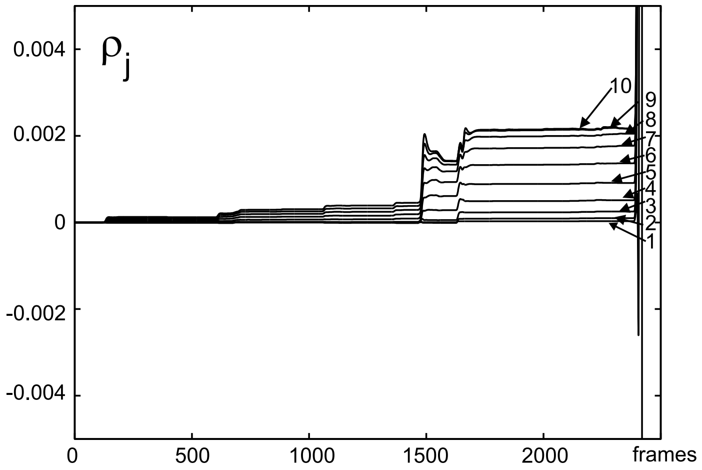

Fig. 5. The weight development of the real robot experiment implementing ISO learning. Parameters: The reflex pathway was set to: $f_0 = 0.01, Q = 0.51$ with a reflex gain of $\rho_0 = 30$. The predictive filters were set to $f_k, k = 0.1/k, k = 1 \ldots 10, Q = 1$ and the learning rate was $\mu = 0.0000035$.

there are important differences. RL is usually implemented as an actor/critic architecture where the critic generates a delta error which tells the actor *what* to do. In other words the delta error is a teaching signal which actively reinforces or penalises actions. However, in ISO3 the signal u_r does not evaluate actions. ISO3 just switches learning on or off but does not force the system towards a certain behaviour. This is an important difference between our ISO3 and RL: The latter uses a global error signal to drive learning which tells the actor *what* to learn whereas our ISO3 tells the actor *when* to learn and leaves the "what" aspect to the actor itself. Learning of the actor in ISO3 is related to spike timing dependent plasticity [11, 12].

Dopamine as a crucial factor for long term potentiation (LTP) has been suggested, for example, in [13, 4] and been reviewed in [14, 15]. Evidence suggests that LTP not only needs coinciding pre- and postsynaptic activity [11, 16] but also dopamine transients as a third factor. Without dopamine no long term potentiation seems to be possible [4]. The third factor of ISO3 can be related to the dopaminergic neurons in the VTA (see Fig. 6) which respond strongly to primary rewards [17]. The VTA in turn is driven by the lateral hypothalamus (LH) which is the primary nucleus which becomes active while eating food. The circuit of LH and VTA could have the task to switch on learning in a number of brain areas like the prefrontal cortex, the hippocampus and the nucleus accumbens which could act as a global signal for learning. In terms of behaviour the nucleus accumbens plays here a central role because it transforms information from the cortex and the hippocampus into motor commands. In our model the learner in Fig 1 can be directly associated with the NAcc: Initially the NAcc

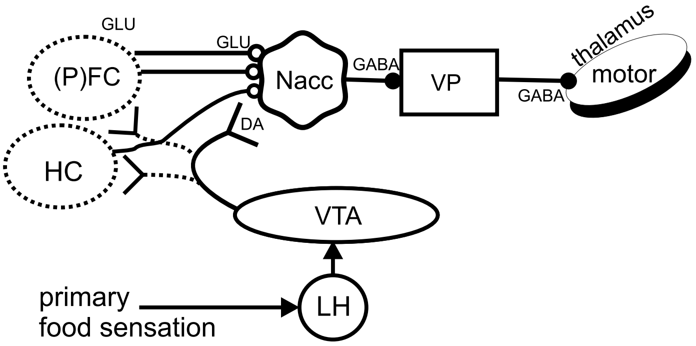

Fig. 6. Simplified diagram of the limbic system. NAcc=Nucleus Accumbens core, HC=Hippocampus, PFC=prefrontal cortex, VP=ventral pallidum, VTA=ventral tegmental area, LH=lateral hypothalamus.

is pre-wired with certain behaviours which are then modified and superseded by learned inputs from the cortex and hippocampus. Thus, learning takes place on top of pre-wired behaviours. Consequently, models like the one developed by Prescott et al. [18] which work with pre-wired behaviour could be upgraded to accommodate learning so that anticipatory behaviour is generated. For the actual learning this means that the dopamine signal does not choose the actions in the striatum but that it rather tells it to learn at a certain moment in time. The striatal neurons would learn locally by themselves with the help of spike timing dependent plasticity and not by a dopaminergic error signal [19]

It is known that dopaminergic activity decreases at the primary reward and builds up at the location of the conditioned stimulus [17]. This behaviour can be re-interpreted if we accept that dopamine is telling the target structure when to learn rather than what to learn: it helps to stabilise behaviour associated with the primary reward because learning is switched off when the signal u_r is no longer happening at the moment the primary reward is experienced. Switching on learning at the first conditioned stimulus preserves the behaviour which is associated with the primary reward.

The applicability of ISO3 learning to other tasks depends on the availability of a third factor. In the food retrieval task it is obvious that the third factor is generated from the contact with food because it is a relevant event. Similarly a relevant event can be found in avoidance tasks, for example collision avoidance. However, the relevance is not tied to the primary trigger, for example food or pain. In more general terms relevance could be derived from novelty which triggers learning when unexpected events have happened which in turn switch on learning to reduce uncertainty.

References

1. Hebb, D.O.: The organization of behavior: A neurophychological study. Wiley-Interscience, New York (1949)
2. Kosco, B.: Differential hebbian learning. In Denker, J.S., ed.: Neural Networks for computing: Snowbird, Utah. Volume 151 of AIP conference proceedings., New York, American Institute of Physics (1986) 277–282
3. Porr, B., Wörgötter, F.: Isotropic Sequence Order learning. Neural Comp. **15** (2003) 831–864
4. Bailey, C.H., Giustetto, M., Huang, Y.Y., Hawkins, R.D., Kandel, E.R.: Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? Nat Rev Neurosci **1**(1) (2000) 11–20
5. Grossberg, S., Schmajuk, N.: Neural dynamics of adaptive timing and temporal discrimination during associative learning. Neural Networks **2** (1989) 79–102
6. Verschure, P.F.M.J., Voegtlin, T., Douglas, R.J.: Environmentally mediated synergy between perception and behaviour in mobile robots. Nature **425** (2003) 620–624
7. Porr, B., Wörgötter, F.: Isotropic sequence order learning in a closed loop behavioural system. Roy. Soc. Phil. Trans. Math., Phys. & Eng. Sciences **361**(1811) (2003) 2225–2244
8. Ziemke, T.: Are robots embodied? In: First international workshop on epigenetic robotics Modeling Cognitive Development in Robotic Systems. Volume 85., Lund (2001)
9. Sutton, R.: Learning to predict by method of temporal differences. Machine Learning **3**(1) (1988) 9–44
10. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. 2002 edn. Bradford Books, MIT Press, Cambridge, MA (1998)
11. Markram, H., Lübke, J., Frotscher, M., Sakman, B.: Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. Science **275** (1997) 213–215
12. Saudargiene, A., Porr, B., Wörgötter, F.: How the shape of pre- and postsynaptic signals can influence stdp: A biophysical model. Neural Comp. **16** (2004) 595–626
13. Centonze, D., Picconi, B., Gubellini, P., Bernardi, G., Calabresi, P.: Dopaminergic control of synaptic plasticity in the dorsal striatum. Eur J Neurosci **13**(6) (2001) 1071–1077
14. Reynolds, J.N., Wickens, J.R.: Dopamine dependent plasticity of corticostriatal synapses. Neural Networks **15** (2002) 507–521
15. Wörgötter, F., Porr, B.: Temporal sequence learning, prediction and control - a review of different models and their relation to biological mechanisms. Neural Comp **17** (2005) 245–319
16. Zhang, L.I., Tao, H.W., Holt, C.E., Harris, W.A., Poo, M.m.: A critical window for cooperation and competition among developing retinotectal synapses. Nature **395** (1998) 37–44
17. Schultz, W.: Dopamine neurons and their role in reward mechanisms. Curr Opin Neurobiol **7**(2) (1997) 191–197
18. Prescott, T.J., González, F.M.M., Gurney, K., Humpries, M.D., Redgrave, P.: A robot model of the basal ganglia: Behaviour and intrinsic processing. Neural Networks, In Press (2006)
19. Dayan, P., Balleine, B.W.: Reward, motivation and reinforcement learning. Neuron **36** (2002) 285–298