4th International Conference on
Development and Learning and on Epigenetic Robotics
October 13-16, 2014. Palazzo Ducale, Genoa, Italy

A Neural Dynamic Model of Associative Two-Process Theory: The Differential Outcomes Effect and Infant Development

Robert Lowe
School of Informatics
Interaction Lab
University of Skövde
Sweden
Email: robert.lowe@his.se

Yulia Sandamirskaya
Theory of Cognitive Systems
Institut für Neuroinformatik
Ruhr-Universität Bochum
Germany
Email: sandayci@rub.de

Erik Billing
School of Informatics
Interaction Lab
University of Skövde
Sweden
Email: erik.billing@his.se

***Abstract*—In animal and human learning, outcome expectancy is understood to control action under a number of learning paradigms. One such paradigm, the differential outcomes effect (DOE), entails faster learning when responses have differential, rather than non-differential, outcomes. The associative two-process theory has provided an increasingly accepted explanation as to how outcome expectancies influence action selection, though it is computationally not well understood. In this paper, we describe a neural-dynamic model of this theory implemented as an Actor-Critic like architecture. The model utilizes expectation-based, or prospective, action control that following differential outcomes training suppresses stimulus-based, or retrospective, action control (known as *overshadowing* in the learning literature). It thereby facilitates learning. The neural-dynamics of the model are evaluated in a simulation of experiments with young children (aged 4-8.6 years) that uses a differential outcomes procedure. We assess development parametrically in neural-dynamic terms.**

I. Introduction

The parallel between animal/human-learning and machine learning based reinforcement learning (RL) has, historically, been strong [1]. However, a phenomenon increasingly recognized in the animal learning literature – the differential outcomes effect (DOE) – has been ignored in machine learning applications of RL. The DOE manifests when training animals on stimulus-response options that have differential outcomes (rewards). It comprises faster learning (than with non-differential outcomes) consequent to using learned outcome expectancies (prospection) to cue correct responses (following presentation of the external stimulus) [2], [3]. Traditional instrumental learning, instead, concerns maintaining the external stimulus in memory (retrospection) for cueing responding. The DOE implies that prospective and retrospective forms of response control can interface through *overshadowing*. Prospection suppresses ('overshadows') the effects of learned retrospection during differential outcomes (DO) training.

Associative two-process theory [2], [3] offers a strong candidate for explaining the DOE. Its focus is on how response (or 'action') control is mediated by both a retrospective route, learned via stimulus-response (S-R) associations, and a prospective route, learned via stimulus-outcome expectancy (S-E) and outcome expectancy-response (E-R) associations. The computational mechanisms for implementing such dual-routes and their mediation, however, have not been identified.

Dynamical systems, and in particular Dynamic Field Theory (DFT) [4], [5] thinking, in the study of infant cognitive capacities and their development have been manifold since the seminal work of Esther Thelen [6]. This paradigm has utilized DFT as a means of modelling the cognitive processes behind infant performance. Thelen et al. produced a mathematical formalism of a motor planning field accounting for the integration of visual stimuli (including location and cue salience), delay time (i.e. memory), and age parameters in a study of motor memory and decision making. In their model, Thelen et al. demonstrated how and why cognitive development may depend intimately on the interrelation among multiple parameters. DFT since then has been successfully applied to modelling development of spatial and visual working memory in infants [7]. This paradigm has as well been used to develop cognitive architectures, which may be linked to physical sensors and motors of embodied agents [8].

In this paper, we present a neural-dynamic-connectionist implementation of associative two-process theory which: a) helps to clarify the computational nature of DO learning using a neural-dynamic perspective, b) can be understood in terms of developmental psychology. The performance of the model is evaluated with respect to a simulations-based replication study of two experiments [9],[10] of infant-learning performance using a DO procedure. The resulting *Actor-DoCritic* (Actor-Differential outcomes Critic) architecture applies to decision making RL (selecting among actions with differential reinforcing outcomes) and faster learning, with potential applications to pedagogics and robotics.

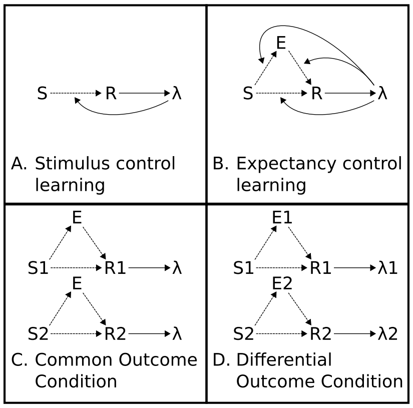
The paper breaks down as follows: In section 2, we present the DOE, associative two-process theory, and a procedure used to investigate both DO and non-DO infant learning performance. In section 3 we present the *Actor-DoCritic*. Section 4 presents results of our simulated replication study of [9],[10]. Section 5 offers concluding comments.

II. Reinforcement learning by differential outcomes

A. *The differential outcomes effect and instrumental learning*

In Instrumental learning theory, contention has historically existed concerning the routes by which animals associate stimuli (S) with responses (R) (cf. [11]). Much debate concerns the nature of S-R versus R-O (outcome) associations. Pertinent to this debate is the Differential Outcomes Effect (DOE).

The DOE concerns improved learning of correct (i.e. rewarded) S-R pairs when *different* rewards are presented for different (correct) S-R pairs. The associative two-process theory of Trapold and Overmier [2], perhaps the leading hypothesis to account for the DOE today [12], [3], builds on classic two-process learning theory [13]. This theory entails separate routes for learning S-R and S-O associations where the latter route has no direct control over responding providing instead a measure of value. Associative two-process theory, to the contrary, proposes that outcome expectancies, during a DO procedure, can cue responses in place of, or in combination with, the external stimuli. The outcome expectancy for a particular reinforcer becomes a stimulus: "the reinforcer itself is part of what is learned" ([12], p.1). In this sense, the classical conception of the stimulus-response-outcome, or (S-R)-O, sequential relation (with S-R in brackets denoting that the stimulus-response association is learned), is more accurately portrayed as (S-E-R)-O where E is the learned expectation tied to a particular outcome. This relationship is captured in Fig. 1.



A. Stimulus control learning

B. Expectancy control learning

C. Common Outcome Condition

D. Differential Outcome Condition

Fig. 1. **The differential outcomes effect**, adapted from [14]. A. S-R associations are simply reinforced. B. Outcome expectation (E) can cue responding following S-E, E-R learning. C. Non-differential outcomes expectations do not provide extra information in response selection. D. Differential outcomes expectations provide additional information to stimuli for cueing responses.

As exemplified in Fig. 1D, under DO training, specific (differential) outcomes expectations (E1, E2) associated with different stimuli are associated with specific responses. This S-E + E-R route under DO training thereby provides more information to such associations than when formed under non-DO training (Fig. 1C). It has been experimentally derived that activity via this S-E + E-R route entails a suppressive effect on conventional S-R output, cf. [14]. It is thereby suggested that internal expectancy cues come to 'overshadow' stimulus cues when concerning response (action) control.

Among others, Pearce [11] suggests that it may not be an easy task to develop a computational model of such an instrumental conditioning approach: "We would need to take account of three different associations that have been shown to be involved in instrumental behavior, S-R, R-US [reinforcement], S-(R-US)" (p.111). From the perspective of (neural) computational modelling the associative two-process theory lends itself to a connectionist approach. However, while S-E, S-R and E-R associations may be developed according to hebbian based learning, two questions are not clearly addressed by the theory (or in the DOE literature):

1) *how should E (expectation) be modelled such that it can differentially cue responses?*
2) *how and when should E-R connections overshadow S-R connections in control of responses?*

From a neurobiological perspective, where do the expectancies come from? [15] has suggested that different structures within the prefrontal cortex capture different dimensions of value used in decision making. [16] has put forward the amygdala as a candidate for learning reward expectancies. [17] suggest that prefrontal cortex (particularly orbitofrontal cortex) may encode reward omission expectation and contributes thereby to the computation of affective working memory (AWM). The activity of the AWM is suggested to suppress that of the more conventional working memory (WM). WM has been linked to the notion of 'retrospection' (keeping in mind a behaviour-eliciting stimulus prior to reward acquisition). AWM, on the other hand, has been linked to the notion of 'prospection' (having in mind the particular reward or goal that can then cue one of a repertoire of behaviours) – see Peterson and Trapold [14]. The suppression of WM by AWM hints as to how, neurobiologically, overshadowing is achieved.

Our modelling approach can be likened to a theoretical behaviourism stance [18] whereby we adopt, where possible, a parsimonious connectionist/neural-dynamic modelling approach. We further frame the model in Actor-Critic terms which helps us address the two above-mentioned questions. The Critic is fundamentally responsible for addressing question (1) while the Actor is responsible for (2). These questions will be further addressed in Section III.

B. *Differential outcomes: learning and development*

Whereas most studies on the DOE have concerned learning and decision making in animals, Maki et al. [9], and later Estevez et al. [10], demonstrated the potential for the DOE to facilitate learning in human infants. Following pre-test and pre-training phases, the experimenters evaluated a further two phases consisting of i) behavioural, and ii) verbal feedback, assessments of the influence of the differential outcomes effect. The latter phase helped to assess the extent prospective control dominated behaviour as infants were asked "What reward am I thinking of?" when presented with the initial stimulus cues. The infants were aged from 4-5.6 years and compared performance on DO and non-DO procedures.

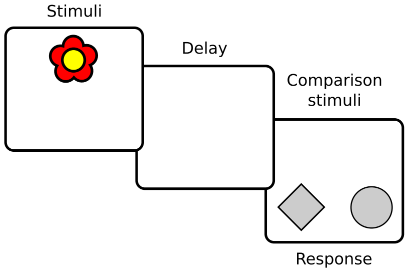


Fig. 2. **Per-trial experimental set-up of Maki et al. (1995) and Estevez et al. (2001)**. The infant must i) observe the sample stimulus, ii) wait during a 2 second delay, iii) point to the correct comparison stimulus – that yields reinforcement. Thus, in this example, S-R1 → reward, S-R2 → no reward.

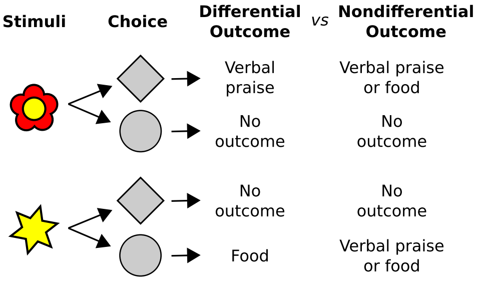


Fig. 3. **Maki et al. (1995)**. Infants are presented one of two stimuli cues per trial associated with comparison stimuli that yield rewarding outcomes. Differential and non-differential outcomes conditions exist. Different symbols are used in Estevez et al. (2001) but the same procedure was applied.

Using the same experimental procedure, Estevez et al. [10] followed up on the work of Maki et al. by evaluating the performance of older infants (4.6 to 8.6 years) to assess whether the DOE would persist developmentally. Figs. 2 and 3 illustrate the experimental procedure followed by Maki et al. and Estevez et al. In Fig. 2 the per-trial presentation is depicted. Firstly, a sample stimulus (image) on the top half of a piece of paper is presented to the infant (and then withdrawn); secondly, a blank piece of paper is presented over a 2 second period (and then withdrawn); thirdly, the comparison stimuli are presented. In the third phase, the infant is required to point (response) to the stimulus that yields the reward. In the DO condition the reward is either food or verbal praise depending on the particular S-R pairing (S1-R1, S2-R2) – see Fig. 3. In the non-DO condition reward – food or verbal praise – is randomly presented. The precise details of the pre-training procedure (not relevant to a computational study) can be found in experiment 1 of Maki et al. [9], and Estevez et al. [10].

III. The Neural Dynamic Actor-DoCritic Model

We now present a continuous time neural-dynamic instantiation of the associative two-process theory (see Fig. 1). Fundamentally, the model is a connectionist implementation of a TD learning architecture. However, the neural dynamic approach to modelling allows for investigating the following:

1) **overshadowing**: how retrospective (WM) and prospective (AWM) memory interact over the intervals between stimulus (S) and outcome (O),
2) **development**: neural field models have been used to develop hypotheses concerning infant development – particularly of WM [19]; here, we are concerned with how retrospective (WM) and prospective (AWM) memory interact over developmental time,
3) **embodiment**: attractor dynamics allow for (noise) robustness and coupling to the sensory and motor systems of real world (e.g. robotics) applications – this element concerns projected future work.

The remainder of this section will describe the rationale and implementation of the neural-dynamic model (Actor-DoCritic), which consists of two parts: 1) *Differential outcomes Critic* (DoCritic), 2) *Differential outcomes Actor*.

A. Differential outcomes Critic

In the DOE, outcome expectations provide internal stimuli that can come to control behavioural responding [12]. The ubiquity of the DOE is testified to by the number of reinforcement dimensions over which its influence has been observed: 1) *type* (e.g. food versus water), 2) *magnitude*, 3) *probability of presentation*, 4) *delay of reinforcement following the stimulus*. This list (cf. [20]) compares to that identified by [21]: "the value of a reward given by an action at a state is a function of reward amount, delay and probability." (p.410).

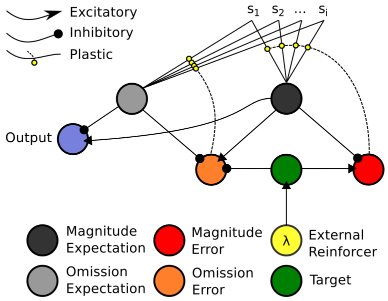


Fig. 4. **A computational model of emotional conditioning**, adapted from [22]. The division into acquisition (magnitude) and omission dimensions allows the model to replicate data from a number of key learning paradigms.

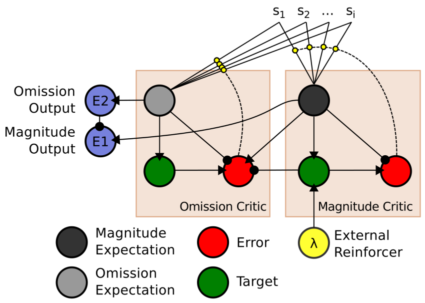


Fig. 5. **The differential outcomes Critic.** TD value expectations for Magnitude and Omission Critics output to relay nodes (expectancy dimensions E1 and E2). The relayed activation of E1 and E2 are used as Actor inputs.

Value functions provide expectations of reinforcing outcomes – providing the 'E' in Fig. 1. The standard bearers of RL in both animal learning and machine learning modelling, i.e. temporal difference (TD) learning [1] and the Rescorla-Wagner model [23], respectively, conflate reinforcement information into a single dimension of value. Therefore, a reinforcer of magnitude 1.0 and presentation probability 0.5 is valued equivalently to one of magnitude 0.5 and presentation probability 1.0. Agents may benefit from multi-dimensional reinforcer information. For example, high magnitude, low probability reinforcers might motivate learning the causal antecedents of the low presentation probability so as to increase future reward yield. In animal learning, the scalar value function has been noted as a key limitation of the Rescorla-Wagner model [24].

Balkenius and Moren [22] present a model of learning (Fig. 4) that addresses the above-mentioned limitation of the Rescorla-Wagner model by deriving a representation of reinforcement omission from a reinforcement acquisition value. Although not explictly noted by the authors, this effectively provides an omission probability as a function of reinforcement magnitude. Fig. 5 depicts our model, referred to as *Differential Outcomes Critic (DoCritic)*. This is an adaptation of the Balkenius model that addresses a further limitation of the Rescorla-Wagner model – delayed discounting of reward [25]. The DoCritic comprises two TD networks for learning magnitude and omission values where omission value is learned as a function of reward (magnitude) value.

The DoCritic calculates value functions that associate presented stimuli with *Magnitude* and *Omission* expectation. We refer to these associations as S_i–$E1$ and S_i–$E2$, respectively, where i is the index of the presented stimulus. The learned value functions produce the standard exponential growth profile of TD discounted delay over the stimulus-outcome interval, where maximum values indicate the expected magnitude and omission probability at reward onset time, in [0,1]. The Magnitude Critic implements standard TD learning though now S_i-$E1$ can be learned but not unlearned. This is consistent with the non-TD implementation of [22].

The Omission Critic's computation is as follows:

- *Precondition for learning*: i) the Omission Critic error node (Fig. 5) updates when reinforcement via the Magnitude Critic 'Target' node is absent at the learned time. Thus, Magnitude Expectation input is no longer 'neutralized'. ii) the omission error is inhibited by 'Target' when reinforcement does arrive at the anticipated time.
- *Asymptotic learning*: i) Omission Expectation inhibits the omission error as a function of S_i-$E2$ weights development. ii) Expectation is learned via the Omission Critic 'Target' in relation to the TD learning discount term γ.
- *Unlearning*: S_i–$E2$ associations decrease as a result of the now unexpected reinforcement input.

Omission Expectation provides output to E2 which inhibits (subtracts from) E1 (Magnitude Expectation output). Equations (1-5) describe the Critic mathematically:

$$V_e(t) = \Sigma_{s \in S}\Big(\theta_{e_s}(t)\phi_s(t)\Big), \tag{1}$$

$$\theta_e(t) = \theta_e(t - \Delta t) + \beta_e \sigma_e \phi_s(t - \Delta t), \tag{2}$$

where $V_e(t)$ is the learned value function (expectation); $\theta_e(t)$ is the value function (S-E) update rule; e in $\{m, o\}$ is an index denoting Magnitude or Omission Critic value functions, respectively; t is time in $[1, T]$ where $T = 100$; s is the number of different stimuli in $[1, S]$ where $S = 2$; β_e is a learning rate in [0,1); Δt is the time window set here to 1; σ_e is the prediction error term; ϕ_s is the 'backward view' (see [1]) eligibility trace of the input stimulus (set to 1 at stimulus onset) calculated as $\phi_s(t) = \phi_s(t - \Delta t)\gamma\lambda_{TD}$, where $\gamma = 1 - \frac{\Delta t}{\tau_m}$ (see below) and $\lambda_{TD} = 1$, implementing TD(1).

$$V_{mrel}(t) = V_m(t) - V_o(t), \tag{3}$$

where $V_{mrel}(t)$ is the magnitude relay output of the Critic. The omission relay $V_{orel}(t) = V_o(t)$ (see Fig. 5).

$$\sigma_m(t) = \lambda(t - \Delta t) + \frac{1}{\Delta t}[(1 - \frac{\Delta t}{\tau_m})V_m(t) - V_m(t - \Delta t)] \tag{4}$$

$$\sigma_o(t) = -\sigma_m(t) + \frac{1}{\Delta t}[(1 - \frac{\Delta t}{\tau_m})V_o(t) - V_o(t - \Delta t)] \tag{5}$$

where σ_m and σ_o represent prediction errors used to update the Magnitude and Omission Critics, respectively, and to approximate them better as Bellman optimality functions; $\lambda(t)$ is the reward signal in [0,1]; τ_m is a time constant.

B. *Differential outcomes Actor*

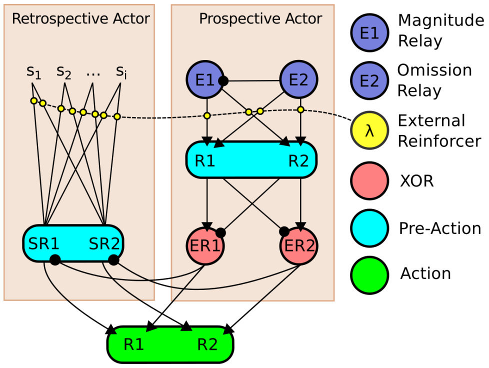


Fig. 6. **The differential outcomes Actor**. The Action field (green) is a one-dimensional (1D) dynamic field over the continuous dimension of 'action orientation'. The Prospective Actor contains 1D fields for each of the $ER1$ and $ER2$ nodes and the Pre-Action field over the same dimension. The Prospective Actor receives expectation/value inputs from the Critic's E1 and E2 nodes. The Retrospective Actor receives stimuli inputs to a 2D Pre-Action field for the continuous stimulus-response dimensions of hue and orientation.

The *DoActor*, depicted in Fig. 6, interfaces with the *DoCritic* (Fig. 5) via inputs from the two Critic networks. These relayed inputs provide calculations of outcome expectation (E) that can be associated through learning with different reinforced responses. The Actor has two networks: *Retrospective Actor* and a *Prospective Actor*. The Actor networks compete for response control. We suggest that this competition should be understood in terms of i) *neural-dynamics*, and ii) *associative learning*.

1) The neural-dynamics of overshadowing: Both Actor networks compute activations across a one-dimensional continuous action space. The Retrospective Actor *Pre-Action* field is a dynamic field, defined over a 2D space that maps a continuous stimulus (hue) dimension to the action space consistent with the spatial nature of the Maki et al./Estevez et al. experimental set-up (see Fig. 2), with hue used as a simplifying visual dimension. The Prospective Actor relays activation from the E_i fields to its *Pre-Action* field and onto its ER_i fields. The two networks use two types of neural-dynamics: 1) Attractor dynamics - in the Retrospective Actor, 2) Exponentially growing activation dynamics - in the Prospective Actor (linearly relayed from the Critic). Prospective Actor output can overshadow Retrospective Actor as a function of i) strength of output, and ii) the comparative strengths of the *amplification constants* of the Actor networks – see below. Eq. 6 describes the activation function of the Retrospective Actor.

$$\begin{aligned} \tau_{SR}\dot{SR}(x,y,t) &= -SR(x,y,t) + h_g \\ &+ \int c_K \omega_c(x'-x, y'-y, \sigma_{g_E}) \Lambda(sr(x',y',t), \beta_{sr}) dx'dy' \\ &- \int c_K \omega_c(x'-x, y'-y, \sigma_{g_I}) \Lambda(sr(x',y',t), \beta_{sr}) dx'dy' \\ &- p_{dev} \cdot c_I + S + sr_{ltm} + c_R \end{aligned} \quad (6)$$

where $\dot{SR}(x,y,t)$ represents the rate of change of the activation level for each node of the two-dimensional (x,y) stimulus(x)-response(y) field as a function of time t. The standard Amari [4] field terms are as follows: τ_{SR} is the time scale of the dynamics; h_g = field resting level; activation in this field is shaped by the local excitation/lateral inhibition interaction profile defined by self-excitatory projections (with amplitude c_K and width σ_{gE}) and inhibitory projections (with strength c_K and width σ_{gI}). The interaction projections are defined by the convolution of a Gaussian kernel with a sigmoidal threshold function. This field also has a number of external inputs: c_I, an inhibition input from the ER_i fields; S, stimuli inputs; sr_{ltm}, a preshape field input updated by reinforcement; c_R, a ridge input for the stimuli. Implementation details of the above can be found at: http://www.cognitionreversed.com/appendices/. c_K and p_{dev} are the amplification constants (independent variables) that determine the balance of the action selection influence of Retrospective and Prospective Actors. The Prospective Actor's fields do not use interaction kernels but instead relay non-transformed activation from the Critic to the Action field.

2) The associative learning of overshadowing: The logic of learning, in this network, is as follows:

- *scaffolding of E-R learning*: i) supra-threshold SR1/SR2 activity inputs to Action layer, ii) reinforcement occurs if a) activation endures (over stimulus-outcome interval), and b) the correct response (R1/R2) is produced, which simultaneously boosts S-R (S_i–SR_i) and S-E associations (E-R associations require S-E associations).
- *E-R learning and overshadowing*: i) E-R associations grow with correct responses; ii) overshadowing occurs as a function of S-E + E-R associative growth and the values of the developmental parameters (see next subsection).

Prospective control, however, is not guaranteed following learning of S-R associations and outcomes. It has been found that following a non-DO procedure, outcome expectancies have no (or minimal) control over responding [3]. We compared these results with those where the DOE is present [12],[3] and noticed that overshadowing requires an XOR mechanism. If an expectation builds up for neither, or both, response options, the retrospective actor (S-R) remains in control. However, where E-R associations uniquely identify only one action, E-R overshadowing of S-R relations results.

3) The developmental parameters of overshadowing:

1) c_K *(retrospective actor)* – this determines the strength of the interaction kernel in the stimulus-response field. It can be tuned to provide one of two types of attractor dynamics on sites of the field: i) *self-stable attractors* – input-dependent supra-threshold activation, ii) *self-sustained attractors* – supra-threshold activation that persists after the stimulus input has been withdrawn.
2) p_{dev} *(prospective actor)* – this determines the strength of a) exponentially growing prospective activation in the Pre-Action field, b) suppression of retrospective control, and c) excitation of the Action field, potentially overriding an existing action/response preference.

Maki et al. [9] found improved learning performance in the DO, compared to the non-DO, condition (see Fig.8(b)) but not for the youngest (4 to 4 years 6 months) of the three age groups studied. Estevez et al. [10], found enhanced performance over age groups with infants in the DO condition outperforming those in the non-DO condition up until the oldest age. The task was considered too simple for the oldest children for a DO effect to be relevant. In our experiment, we have two independent variables (IVs): c_K and p_{dev} the values of which we propose constitute a developmental trajectory.

IV. Results

As a necessary first test of the model we assessed the DoCritic on learning acquisition, extinction and reacquisition profiles. Realistic animal/human learning models are required to capture such fundamental empirically derived learning data [24]. Following this, we show that the same model can capture data from the more complex DO and non-DO procedures.

A. Validation of the DoCritic

As an illustration of the theoretical relevance of the DoCritic, we evaluated the Actor-DoCritic on the *acquisition-extinction-reacquisition* learning paradigm. The model of [22] produces the desired profile for this learning paradigm but without using delay discounted value functions to bridge the stimulus-outcome interval. Our model utilizes a stimulus-outcome interval of 25 time steps with values in Fig. 7 plotted at per-trial reward onset and over 100 trials.

In the acquisition phase, following the presentation of a single stimulus (S) and presentation of reward following the response, response rate climbs at a negatively accelerated rate to asymptote over learning trials. During the response extinction phase the Omission Critic develops an omission probability expectation (S-E2). This inhibits (is subtracted

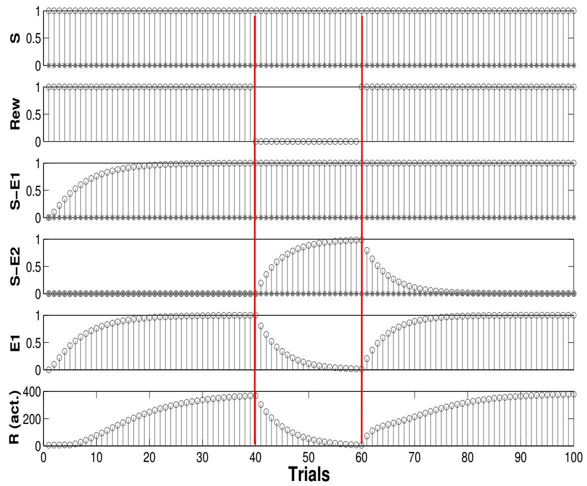


Fig. 7. **The reacquisition effect.** The red vertical lines demarcate three phases of learning: Phase 1 = S-R acquisition, Phase 2 = S-R extinction, Phase 3 = S-R reacquisition. Phase 3 response rate is higher, earlier, than that of phase 1 in spite of response exctinction at the end of phase 2, demonstrating the importance of S-R associative *histories* [18] to response performance.

from) the output of the Magnitude Critic (S-E1) leading to response extinction, at a negatively accelerated rate. In the reacquistion phase, response rate recovers quickly relative to phase 1 as the Magnitude Critic value (Magnitude Expectation) does not decay during the extinction phase and the Omission Expectation update rate (β_e in Eq. 2) is greater than that of the Magnitude Critic. This ensures a smaller difference in *S-E1 minus S-E2* in early trials of phase 3 than in early trials of phase 1. The delayed acquisition response rate relative to the E1 (Magnitude Critic) output is caused by i) different time constants τ for the Prospective Actor fields through which the Critic activity is relayed, and ii) the precondition of E-R associative learning. The DoCritic here is thus able to qualitatively reproduce the *savings effect* where the single stimulus (S) and response (R) apply. The same ('Critic') model can be used to explain data generated on differential outcomes reinforcement schedules. In the example shown in Fig. 7, we have used the parameterization for the most 'developed' network – independent variable (IV) = 6, see next subsection.

B. Validation of the Actor-DoCritic: Developmental studies

We sought to replicate, in simulation, the cumulative results of the Maki et al. [9] and Estevez et al. [10] experiments. We used 6 values of the independent variable (IV) across two conditions (DO vs non-DO), where [p_{dev}, c_K] pairs in {0.0 14.35; 5.20 14.53; 8.24 15.00; 10.4 16.30; 12.07 19.82; 13.44 29.38}. These values (to 2 d.p.) were obtained according to:

$$p_{dev}(n) = log(n) * p_C \tag{7}$$

$$c_K(n) = r_C1 * (0.95 + r_C2 * e^n) \tag{8}$$

where n is the IV number in (1,6); p_C = 7.5; r_C1 = 15; r_C2 = 0.0025. These functions were chosen with the assumption of a linear relation between developmental parameter growth and performance (correct response selection) following observable natural logarithmic and exponential functions in Fig. 8(b). This tenuous assumption was made in the absence of existing hypotheses. We also carried out *control* runs where IV values used linear growth functions with lower and upper ranges dictated by the *experimental* growth functions. In the control, no observable difference occurred between DO and non-DO conditions (see http://www.cognitionreversed.com/appendices/).

In the Maki et al. and Estevez et al. investigations 32 trials were used in each experiment for each age group in each condition (DO vs non-DO) with approximately 7 children per age group. In our simulation, we ran 20 independent runs ('subjects') over 60 trials in each condition. The longer trial length reflects use of omission probability as the DO dimension in our study. Thus, many trials were needed to grow 'value' in the Omission Critic to reflect omission probability. In the DO condition:

- S1-R1→0.0 omission prob., magnitude=1,
- S2-R2→0.8 omission prob., magnitude=1,

In the non-DO condition:

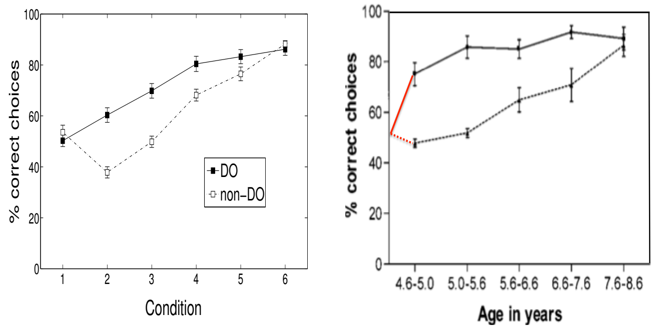
- S1-R1→0.4 omission prob., magnitude=1,
- S2-R2→0.4 omission prob., magnitude=1,

In neither condition were alternative S-R pairs reinforced. In the Maki et al. and Estevez et al. investigations a 2 second long stimulus - response option delay existed (trace conditioning) where outcome immediately followed correct response. In our simulation, the network was randomly presented (no stimulus was presented more than three consecutive times) with one of the two stimuli at t=25 which persisted for 25 steps. The λ input was presented at t=75.

1) Model performance: In Fig. 8(a) is shown the mean percentage of correct responses made by the *Actor-DoCritic*. The model replicates the findings of i) no significant difference (SEM bars do not overlap – $p < 0.05$) in DO vs non-DO performance in IV1 ('youngest') and IV6 'oldest' ages, ii) chance performance (i.e. $\mu < 0.6$) in non-DO performance in IV1-IV3, iii) above chance performance in DO for IV3-IV6, iv) above chance performance in non-DO for IV4-IV6. In the Maki et al. and Estevez et al. experiments significant main effects, using analysis of variance (ANOVA) were found between DO and non-DO conditions but, and consistent with our findings, independent t-tests showed no significant difference at the youngest (our IV1) and oldest (our IV6) ages.

We now evaluate this result in terms of associative two-process theory and the neural-dynamics of overshadowing.

2) Associative two-process learning: Fig. 9 shows associative two-process learning performance for the final run of the network over all trials in the most 'mature' parameterization (IV = 6) in the DO condition. The subplots from top to bottom, showing the associative paths from stimulus (S) to response (R - action), are as follows (see Fig. 6 for reference): 1) S - stimulus input (1 or 2), 2) $S - R1$ shows $S - R$ field preshape values (sr_{ltm}) for response 1 and for S in $\{1, 2\}$, 3) $S - R2$ shows $S - R$ field preshape values for response 2 and for S in $\{1, 2\}$, 4) $S - E1$ shows the growth of the E1 (magnitude)



(a) **Mean % correct response.** (b) **Maki + Estevez data.**

Fig. 8. **Mean performance over trials.** (a) Simulation results for IV values 1-6 in DO and non-DO conditions; (b) % correct response on the Estevez et al. investigation over five age groups, the Maki et al. data for infants aged 4 to 4.6 years is superimposed in red – chance performance; solid plots = DO condition, dashed plots = non-DO condition. The profile in (a) was not replicated when using a linear growth function for p_{dev} and c_K – overlapping error bars between DO and non-DO conditions for 5/6 IVs (20 runs).

expectation for each stimulus, 5) $S-E2$ shows the growth of the E2 (omission) expectation for each stimulus, 6) $E1-R$ connects E1 and R (pre-act) in $\{1,2\}$, 7) $E2-R$ connects E2 to R (pre-act) in $\{1,2\}$, 8) R shows response selection in the action field, 9) Corr. shows whether the response is correct. In plots 2-3, $S=1$ is depicted in dark grey and $S=2$ in light grey. In plots 6-8, R1 is depicted in purple and R2 in green.

In Fig. 9, we see a typical learning trajectory of the Actor-DoCritic: 1) S-R connections form allowing for activation in the SR field to bridge the stimulus-outcome interval, 2) reinforced correct responses simultaneously increase S-E associations (note, S-E2 requires more trials for an accurate omission probability representation), 3) learning of S-E (pre-synaptic values) is a prerequisite to E-R association formation, 4) E-R growth leads to stronger outputs from the ER1/ER2 fields to suppress SR activation, 5) since all weights in the model have a decay rate, suppressed SR activity leads to decay of S-R weights as sub-threshold activation is not reinforced. In this sense point (1) constitutes a scaffolding of prospective (E-R) learning via initial retrospective (S-R) learning. The same effect is not observed in the non-DO condition (for IV=6) as when E-R weights develop (slowly for E2-R) multiple inputs result (via the Pre-Action field) to ER1 and ER2 fields and the effective XOR mechanism prevents any overshadowing. Thus, S-R weights do not decay in this control condition. Nevertheless, the strong *self-sustained* attractors in the SR field (due to high c_K) permit learning of correct responding via cancelling out of the effects of noise in the action field. E2 (omission)-R connections decay in the absence of reinforcement, and 'hike' up when reinforcement arrives. S-E2 shows the opposite effect – when reinforcement arrives, omission probability approximation drops. It can be seen that the probability is approximate and takes time to grow.

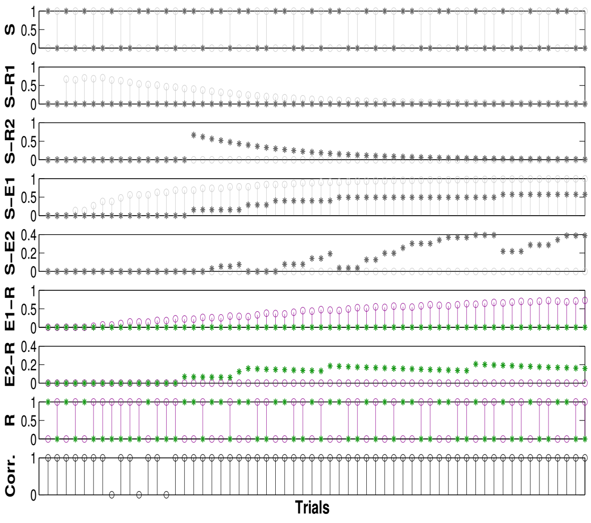


Fig. 9. **DO associative two-process learning –** $IV=6$

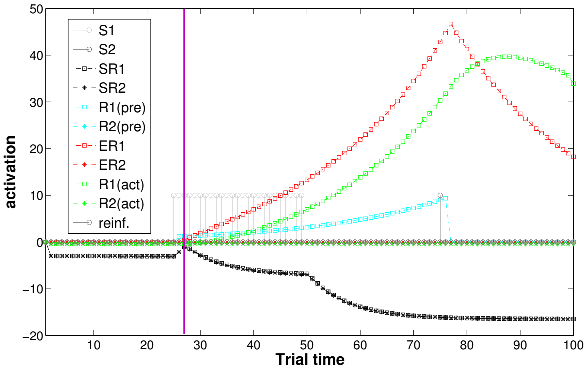


Fig. 10. **DO trial neural-dynamics –** $IV=6$. The stemmed lines at $t=25-50$ indicate stimulus presence (scaled x10). The stemmed line at $t=75$ indicates the reinforcer presence following a correct response. The vertical purple line points at when ER1 overshadows SR1 response control.

3) Overshadowing: Figs. 10 and 11 show the final trial for the final run in IV = 6 for DO and non-DO conditions. The figures depict trial neural-dynamics as retrospective actor and prospective actor compete for response control. Fig. 10 shows a case of prospective overshadowing where $ER1$ activation dominates in reference to an $S1-R1$ pairing. This occurs following the learning of $E-R$ associations (Fig. 9). At the extreme, ER activations have the power to change the action choice previously promoted by the $S-R$ field ('dithering'). In Fig. 11 no overshadowing occurs as ER fields are deactivated by the XOR mechanism. This owes to the learning of two $E-R$ associations leading to two supra-threshold outputs at the Pre-Action field – R1(pre) and R2(pre) – and thus mutual inhibition in the ER1 and ER2 fields. A *self-sustained* attractor in the SR field, however, ensures the stimulus-outcome interval is bridged where R2 (act) is the reinforced response.

We tested for tendency to change action choice ('dither' between supra-threshold activations at the two action sites in the action field) between the period of CS onset and final action (at US onset). We looked only at performance in the second half of all runs over IV2-6 (i.e. from trial 31-60). We discounted IV=1 since correct choice performance in both DO and non-DO conditions was at chance levels. A one-way ANOVA found a main effect at the 0.01 level of significance ($p=0.0026$) with the DO condition producing a lower dithering mean (1.526) than for the non-DO condition

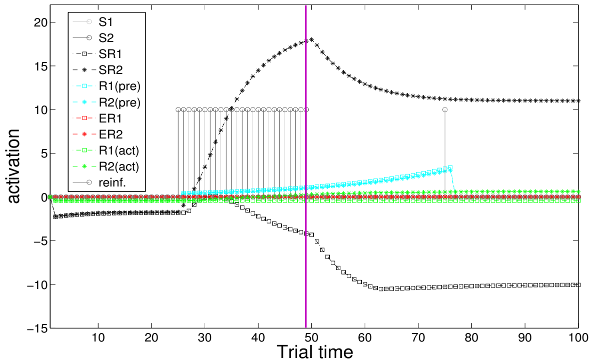


Fig. 11. **non-DO trial neural-dynamics –** $IV = 6$. The purple line shows the point at which S2 is withdrawn and SR2 falls into a *self-sustained* attractor.

(2.746). This phenomenon owes to the strong ER field output (as exemplified in Fig. 10) inducing more 'decisiveness'.

V. Conclusion

In this paper we presented a neural-dynamic model of associative-two process theory via an Actor-Critic architecture. We demonstrated how: 1) outcome expectancy dimensions can be modelled, 2) prospective overshadowing of retrospective response control can occur neural-dynamically, over learning and development. This has allowed us to understand how an associative two-process account may explain, computationally, mediation of prospective and retrospective response control.

Neural-dynamic models have previously been used to assess the development of spatial working memory, SWM (similar to our model's restrospective memory). It has been suggested [6] that SWM develops to effectively compete with another type of memory (motor memory) leading to more flexible response control. The implication of our model is that prospective (affective working) memory (AWM) overshadows retrospective memory (SWM) in early development. The developmental profile of restrospective and prospective memory, as captured by our model parameterization of development, indicates that SWM develops at an *initially slower pace* compared to AWM.

Our simulations offer an empirically testable hypothesis of neural-dynamic overshadowing: *dithering* – changeable action tendencies – should be lower under DO than non-DO conditions. In our model, dithering is a symptom of the faster learning in the DO condition, i.e. the Prospective Actor produces stronger output than the Retrospective Actor within this developmental range. Finally, we suggest that this model has particular application to online decision making in robotics scenarios via utilizing differential DO expectancy information.

VI. Appendix

Supplementary material (model and results) can be found at: http://www.cognitionreversed.com/appendices/.

Acknowledgment

This work was funded through the 7th framework of the EU in grant #270247 (NeuralDynamics project).

References

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

[2] M. Trapold and J. Overmier, "The second learning process in instrumental learning." in *Classical conditioning 2: Current research and theory*, A. H. Black and W. F. Prokasy, Eds. New-York: Appleton-Century-Crofts, 1972, pp. 427–452.

[3] J. M. Holden and J. B. Overmier, "Performance under differential outcomes: Contributions of Reward-Specific Expectancies," *Learning and Motivation*, pp. 1–14, 2014.

[4] S. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological Cybernetics*, vol. 27, pp. 77–87, 1977.

[5] G. Schöner, *Cambridge handbook of computational cognitive modeling*. Cambridge, UK: Cambridge University Press, 2008, ch. Dynamical systems approaches to cognition, pp. 101–126.

[6] E. Thelen, G. Schöner, C. Scheier, and L. B. Smith, "The dynamics of embodiment: a field theory of infant perseverative reaching." *The Behavioral and brain sciences*, vol. 24, no. 1, pp. 1–34, 2001.

[7] J. P. Spencer and G. Schöner, "Bridging the representational gap in the dynamic systems approach to development," *Developmental Science*, vol. 412, pp. 392–412, 2003.

[8] Y. Sandamirskaya, S. Zibner, S. Schneegans, and G. Schöner, "Using Dynamic Field Theory to extend the embodiment stance toward higher cognition," *New Ideas in Psych.*, vol. 31, no. 3, pp. 322–339, 2013.

[9] P. Maki, J. B. Overmier, S. Delos, and A. J. Gumann, "Expectancies as Factors Influencing Conditional Discrimination Performance of Children," *The Psychological . . .* , vol. 45, pp. 45—-71, 1995.

[10] A. Estevez and et al., "The Differential Outcome Effect as a Useful Tool to Improve Conditional Discrimination Learning in Children," *Learning and Motivation*, vol. 32, no. 1, pp. 48–64, Feb. 2001.

[11] J. Pearce, *Animal Learning and Cognition: An Introduction, 3rd Edition*. Psychology Group, Taylor and Francis., 2006.

[12] P. J. Urcuioli, "Behavioral and associative effects of differential outcomes in discrimination learning," *Animal Learning & Behavior*, vol. 33, no. 1, pp. 1–21, Feb. 2005.

[13] O. H. Mowrer, "On the dual nature of learning a reinterpretation of "conditioning" and "problem-solving"." *Harvard Educational Review*, vol. 17, pp. 102–148, 1947.

[14] G. B. Peterson and M. A. Trapold, "Effects of altering outcome expectancies on pigeons' delayed conditional discrimination performance," *Learning and Motivation*, vol. 11, no. 3, pp. 267–288, Aug. 1980.

[15] S. W. Kennerley and M. E. Walton, "Decision making and reward in frontal cortex: complementary evidence from neurophysiological and neuropsychological studies." *Behavioral neuroscience*, vol. 125, no. 3, pp. 297–317, 2011.

[16] L. M. Savage and R. L. Ramos, "Reward expectation alters learning and memory: The impact of amygdala on appetitive-driven behaviors," *Behavioural Brain Research*, vol. 198, pp. 1–12, 2009.

[17] M. Watanabe and et al., "Reward Expectancy-Related Prefrontal Neuronal Activities: Are They Neural Substrates of 'Affective' Working Memory?" *Cortex*, vol. 43, pp. 53–64, 2007.

[18] J. Staddon, *The New Behaviorism.* Psychology Press, 2014.

[19] J. S. Johnson, J. P. Spencer, and G. Schöner, "Moving to higher ground: The dynamic field theory and the dynamics of visual cognition," *New Ideas in Psychology*, vol. 26, pp. 227–251, 2008.

[20] A. Friedrich and T. Zentall, "A differential-outcome effect in pigeons using spatial hedonically non-differential outcomes," *Learning and Behavior*, vol. 39, pp. 68–78, 2011.

[21] K. Doya, "Modulators of Decision Making," *Nature Neuroscience*, vol. 11, no. 4, pp. 410–416, 2008.

[22] C. Balkenius and J. Morén, "Emotional Learning: A Computational Model of the Amygdala," *Cybernetics and Systems: An International Journal*, vol. 32, pp. 611–636, 2001.

[23] R. A. Rescorla and A. R. Wagner, *Classical Conditioning II: Current Research and Theory.* New York: Appleton- Century-Crofts, 1972, ch. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement.

[24] R. Miller, R. Barnet, and N. Grahame, "Assessment of the Rescorla-Wagner Model," *Psych. Bulletin*, vol. 117, no. 3, pp. 363–386, 1993.

[25] Y. Niv, "Reinforcement learning in the brain'." *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139–154, 2009.