Learning the condition of satisfaction of an elementary behavior in dynamic field theory

Luciw, M; Kazerounian, S; Lahkman, K; Richter, M; Sandamirskaya, Y

Abstract: In order to proceed along an action sequence, an autonomous agent has to recognize that the intended final condition of the previous action has been achieved. In previous work, we have shown how a sequence of actions can be generated by an embodied agent using a neural-dynamic architecture for behavioral organization, in which each action has an intention and condition of satisfaction. These components are represented by dynamic neural fields, and are coupled to motors and sensors of the robotic agent.Here,we demonstratehowthemappings between intended actions and their resulting conditions may be learned, rather than pre-wired.We use reward-gated associative learning, in which, over many instances of externally validated goal achievement, the conditions that are expected to result with goal achievement are learned. After learning, the external reward is not needed to recognize that the expected outcome has been achieved. This method was implemented, using dynamic neural fields, and tested on a real-world E-Puck mobile robot and a simulated NAO humanoid robot.

DOI: 10.1515/pjbr-2015-0011

Originally published at:
Luciw, M; Kazerounian, S; Lahkman, K; Richter, M; Sandamirskaya, Y (2015). Learning the condition of satisfaction of an elementary behavior in dynamic field theory. Paladyn, 6(1):180-190. DOI: 10.1515/pjbr-2015-0011

Research Article Open Access

Matthew Luciw*, Sohrob Kazerounian*, Konstantin Lahkman, Mathis Richter, and Yulia Sandamirskaya

Learning the Condition of Satisfaction of an Elementary Behavior in Dynamic Field Theory

DOI 10.1515/pjbr-2015-0011
Received February 8, 2014; accepted July 29, 2015

Abstract: In order to proceed along an action sequence, an autonomous agent has to recognize that the intended final condition of the previous action has been achieved. In previous work, we have shown how a sequence of actions can be generated by an embodied agent using a neural-dynamic architecture for behavioral organization, in which each action has an *intention* and *condition of satisfaction*. These components are represented by dynamic neural fields, and are coupled to motors and sensors of the robotic agent. Here, we demonstrate how the mappings between intended actions and their resulting conditions may be learned, rather than pre-wired. We use reward-gated associative learning, in which, over many instances of externally validated goal achievement, the conditions that are expected to result with goal achievement are learned. After learning, the external reward is not needed to recognize that the expected outcome has been achieved. This method was implemented, using dynamic neural fields, and tested on a real-world E-Puck mobile robot and a simulated NAO humanoid robot.

Keywords: Neural Dynamics; Cognitive Robotics; Behavioral Organization

1 Introduction

Recently, some of us have introduced a computational neural-dynamic model of *intentional actions*, based on Dynamic Neural Fields [15, 16]. Intentional actions are named as such due to Searle's theory of intentionality [24]. Two control components are essential for any intentional action: a perceptual representation of the intention itself, which must also be useful for guiding the motor system (such as the foveal location of a target object), and a representation of the *condition of satisfaction* useful for signaling that the objective, or goal, of the action has been successfully achieved.

In our model, intentional actions are represented in the neural-dynamic controller by *elementary behaviours* (EBs), each consisting of a neural-dynamic realization of intention and condition of satisfaction (CoS). The framework of Dynamic Neural Fields (DNFs) was used to implement the intention and CoS as attractor dynamics, defined over continuous parameter spaces [17, 19, 21], and coupled to the sensory and motor systems of the embodied agent. The intention DNF represents the sensorimotor parameters of the current action and directs the agent's attentional shifts and movements. The CoS DNF receives perceptual input and is activated when this input overlaps with an internal bias, projected from the intention DNF, which specifies the desired final state of the action. Complex actions require coordination between a number of simpler EBs, such that each EB is activated in the appropriate order, persists as long as necessary in order to achieve its behavioral goal, and is ultimately deactivated once the goal is achieved.

We have previously demonstrated how sequences of goal-directed actions may be generated in this neuraldynamic framework for behavioral organization by linking the neural-dynamic architecture to sensors and motors of a humanoid robot [15, 16]. We have also demonstrated how sequences of EBs may be learned from delayed rewards by combining the neural-dynamic architecture with reinforcement learning [26], by making use of eligibility traces that are implemented as neural dynamic item-and-order working memory [12]. In that prior work, the structure of an EB – i.e., the coupling between the intention and the CoS DNFs that encodes the anticipated outcome of an action – was pre-wired during design of the neural-dynamic architecture. For instance, the intention of the EB "search for color" encoded the color of the object, at which the

***Corresponding Author: Matthew Luciw, Sohrob Kazerounian:** The Swiss AI Lab IDSIA, USI & SUPSI, Galleria 2, 6900 MannoLugano, Switzerland, E-mail: matthew@idsia.ch
Konstantin Lahkman: NBIC-Centre, National Research Center, Kurchatov Institute, Moscow, Russia
Mathis Richter, Yulia Sandamirskaya: Ruhr-Universität Bochum, Institut für Neuroinformatik, Bochum, Germany

robot's gaze should be directed. The connection weights between the intention and the CoS DNFs of this EB were chosen such that the CoS DNF was biased to be sensitive to this color, present in the central portion of the camera image. In the present article, we explore how this link from an active intention to a CoS may be learned autonomously by a reward-driven associative learning process, fully within the neural dynamics framework. We demonstrate the functioning of the developed neural-dynamic architecture for learning conditions of satisfaction in two example scenarios with embodied robotic agents: a simulated NAO robot and a physical E-Puck robot.

2 Dynamic Field Theory

2.1 Overview

Dynamic Field Theory originates in analysis of the activation dynamics of neuronal populations. Activation of such neuronal populations during a perceptual or motor task can be modelled by a neural field, which assumes homogeneous connectivity among neurons in the population and averages away the discreetness of individual neurons and the spiking nature of their activation. Amari [1], Wilson and Cowan [29], and Grossberg [9] were among the first to mathematically formalise the activation of a neuronal population as a Dynamic Neural Field (DNF) equation:

$$\tau\dot{u}(x,t) = -u(x,t) + h_u + Sfu(x^{'},t)\omega(x^{'}-x)dx^{'} + I_t(x,t). \tag{1}$$

Here, the activation of a DNF is denoted by $u(x,t)$, where x is the parameter that spans the dimension over which the DNF is defined – i.e. a behavioural dimension, to which the neurons in the modelled population are sensitive. t is time, τ is the time-constant of the dynamics that determines how fast the activation converges towards the attractor, defined by the three last terms on the right hand-side of the equation: the negative resting level h_u, the homogeneous lateral interactions, shaped by the interaction kernel ω, typically a sum of Gaussians with a narrow positive part and a broader, but weaker negative part ("local excitation, global inhibition" or "Mexican hat" kernel) and by the output non-linearity of the DNF, $f[\cdot]$, typically a sigmoid; the last term of the equation is external input, which drives the DNF and comes either from another DNF (neuronal population) or a sensory system.

Lateral interactions of a DNF ensure the existence of a localised activity bump as a stable solution of the dynamics, described by Eq. 1: in response to a distributed, noisy input, a DNF builds a localised bump of positive activation, which is stabilised against decay by the positive part of the interaction kernel and against spread by its negative part. These localised activity bumps, or peaks, are units of representation in Dynamic Field Theory of Embodied Cognition [21], in which DNFs are used to model behavioural signatures of perceptual and motor decision making, working memory, category formation, attention, recognition, and learning [11, 17, 25]. DNF architectures of various cognitive functions were used to both model human behavioural data and to control autonomous robots, in order to demonstrate that the architectures may indeed be embodied and situated [7, 19].

The ability of Dynamic Neural Fields to form and stabilize robust categorical outputs from noisy, dynamical, and continuous real-world input are the basis for their use in the sensorimotor interfaces of cognitive systems, including cognitive robots [3]. DFT has been applied across a number of domains in robotics, from low-level navigation dynamics with target acquisition based on vision [2], object representation, dynamic scene memory, and spatial language [19] to sequence generation and sequence learning [12, 18].

These activation peaks in DNFs represent perceptual objects or motor goals in the DFT framework. Multiple coupled DNFs spanning different perceptual and motor dimensions can be composed into complex DNF architectures to organize robotic or model human behavior. A single DNF builds a stable localised peak that may track the sensory input. In order to generate a sequence of behaviours, an additional mechanism is needed, which allows this attractor solution to be destabilised when the behavioural goal of the current action is achieved. This led to development of the building block of DNF architectures for behavioural organisation – an Elementary Behavior that ensures that dynamical attractors are stabilised and destabilised as the agent proceeds from one behaviour to the next one. We present these building block next.

2.2 Elementary Behaviors

An elementary behavior in DFT (Fig. 1; [16]) consists of intention and condition of satisfaction DNFs. An intention DNF either primes the perceptual system of the agent (e.g. to cue it to be more sensitive to a particular feature) or drives the motor dynamics of the agent directly (e.g. setting attractors for the motor dynamics). The CoS DNF in turn receives a top-down bias from the intention DNF that specifies which perceptual inputs are signalling the successful completion of the intended action. To enable this, two inputs converge on the CoS DNF: one from the intention DNF and one from a perceptual DNF, which is connected to a sensor and builds activity peaks over salient portions of the sensory stream. If the two inputs match in the dimension of the CoS DNF, an activity peak emerges in this field, inhibiting the intention DNF of the EB. The intention DNF follows the generic DNF equation, Eq. (1).

Equation 2 describes our dynamics for a CoS DNF:

$$\tau \dot{v}(y,t) = -v(y,t) + h_v + R(t) + Sfv(y', t)\omega(y' - y)dy' \\ + Sm[W(x,y,t)]f[u(x,t)]dx + I_{sens}(y,t). \quad (2)$$

Here, $v(y, t)$ is activation of the CoS DNF, where y is the parameter which corresponds to a perceptual feature to which the CoS DNF is sensitive. $I_{sens}(y, t)$ is the sensory input that comes from a perceptual DNF, which in turn is directly coupled to the agent's sensors. $R(t)$ is the reward signal, which provides a global boost to the CoS field when an internal drive is satisfied. $W(x, y, t)$ is the two-dimensional weight function that projects positive activation of the intention DNF onto CoS DNF. Learning dynamics for this weight function are described in Section 3.

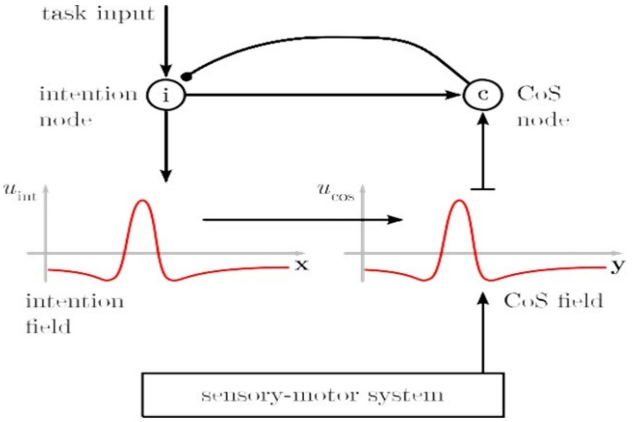

Figure 1: Schematic representation of a generic elementary behavior.

The intention and CoS DNFs are associated with intention and CoS *nodes*, respectively. These nodes facilitate the sequential organization of EBs. While the DNFs are relevant for intra-behavior dynamics, such as selection of the appropriate perceptual inputs for a given behavior, the nodes play a role on the level of inter-behavior dynamics (i.e., switching between behaviors). In previous work, we have shown how EBs may be chained according to rules of behavioral organization [15, 16], serial order [5, 6, 18], or the value-function of a goal-directed representation [12].

Super-threshold activation of the condition of satisfaction DNF generates a signal, which denotes that the intention of its EB is successfully achieved. For instance, the CoS DNF for the behavior 'find the red object' would detect when a large red object is present in the visual field. Activation of the CoS is determined both by the particular dimension(s) of the given CoS field, as well as the synaptic connection weights from the intention field to the CoS field. While the dimensions of the field reflect which sensory dimensions the robot is sensitive to, the weights shape the preactivation in the CoS field and make specific regions of the field sensitive to perceptual input. This can be thought of as an anticipatory attentional bias.

In our previous work, the intention to CoS weights W (see Eq. 2) were 'hardcoded' into the architecture. The dimensions of the CoS field and the synaptic weights converging onto the field were designed such that they would produce super-threshold CoS activation (i.e., a peak in the CoS field) under the desired conditions. Although such hardcoded constraints have successfully been shown to generate desired behaviors in robotic agents (see e.g., [15]), we next address the question of how the structure of an EB can be learned without a priori design of the intention to CoS coupling.

3 Learning a Condition of Satisfaction

Here, we present a DFT mechanism for learning a condition of satisfaction through reward-gated associative learning. The basic Elementary Behavior is augmented with adaptive weights from the intention field to the CoS field. The learning rule tunes the weights when a reward signal is received, increasing weights that connect to the CoS DNF's features that are present in the stimuli, and decreasing weights to the locations of the CoS DNF that correspond to features that are not present. Features could correspond to many different characteristics of the environment, depending on the robot and the desired behavior. One of the simplest features is color (which is what we use in our experiments). The learned values of the weights ultimately specify which perceptual features were most often associated with reward. After learning, the function of the weights is to boost the CoS field locally, by priming the features which were learned to be associated with reward. Once those features are perceived, the activity of the CoS field reaches threshold, signalling that the active behaviour has achieved its goal, at which point the reward signal driven by an internal drive is not needed.

In the present work, the reward signal is designed to come from a teacher, who could be training the robot

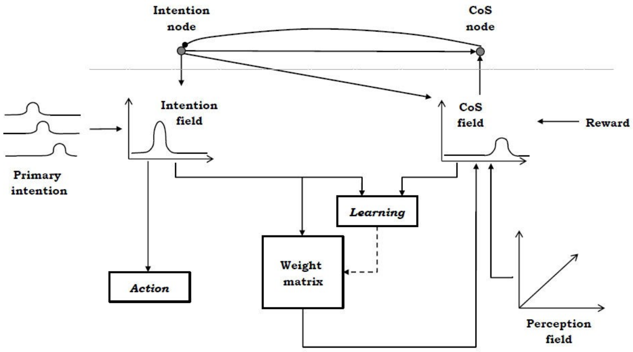

Figure 2: Architecture for CoS learning.

how to complete its elementary behaviors. This is similar in spirit to work involving the SAIL robot, which was trained to perform obstacle avoidance in real time by reward and punishment signals coming from following a teacher's proper and timely usage of "good" and "bad" buttons [28].

An alternate interpretation that doesn't require a teacher is that the rewarding signal is associated with innate internal drives. As mentioned, these drives can be similar to the prototypical drives suggested by Woodworth, e.g. hunger and thirst [30]. Drives such as these serve as internal forces that initiate behaviors and agents are rewarded when the drives are satisfied [10].

The behaviors learned in order to satisfy these drives can be internalized, and recalled, in circumstances similar to those involving drive satisfaction, but where there is no actual (external) satisfaction (reward signal). Even though the agent does not achieve actual immediate reward of the type that satisfies the primitive internal drive that caused the behavior to be formed, it may find the behavior useful in another context, perhaps in combination with other behaviors, to reach an alternate source of reward.

3.1 Reward Gated Associative Learning in Dynamic Fields

The DFT learning process leads to the formation of memory traces in the mapping between the intention and CoS dynamic neural fields. Fig. 2 illustrates a sketch of the learning architecture.

There are two dynamic neural fields, for intention and CoS, respectively, each following Equation (1) and Equation (2), respectively. The intention DNF builds activity peaks with different locations in the field's dimension depending on the currently active internal drive (primary intention) and activates the agent's behavior (action). The CoS field receives input from the perception DNF and input from the Intention DNF through a weight matrix.

The reward signal, $R(t)$ in Eq. (2), provides a global boost to the CoS field, with the purpose of pushing perceptually induced activations above the output threshold, to enable learning of weights between the active regions of the intention and CoS DNFs. We conceptualized the reward signal as binary ($R(t) \in \{0,1\}$).

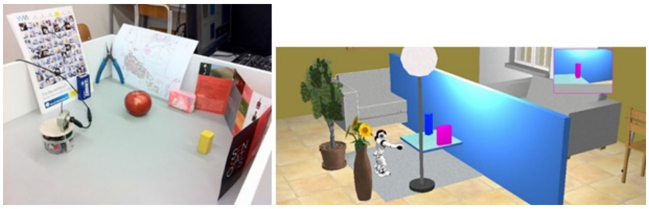

Figure 3: Experiment environments for E-Puck (Left) and NAO (Right).

The two-dimensional weight function, $W(x, y, t)$, maps the output of the intention DNF onto the CoS DNF, as shown in Fig. 5. $W(x, y, t)$ is updated according to the reward-driven learning rule:

$$\tau_l \dot{W}(x, y, t) = \lambda R(t) - W(x, y, t) + fv(y, t) \cdot fu(x, t) \qquad (3)$$

Note that the weights are only updated when a nonzero reward signal $R(t)$ is perceived. The intention field output $f[u(x, t)]$ also gates the learning, such that weight values can only be updated along the "ridge" of $W(x, y)$ selected by intention field peak location x. For weights without support from the CoS field $f[v(y, t)]$, their values will decay according to $-W(x, y, t)$. The weights with perceptual support have their values increased. λ is a learning rate parameter.

Fig. 5 shows an example of a mapping between two, one-dimensional, intention and CoS dynamic neural fields. In this case, the coupling between them is 2D, and can be visualized easily. The effects of the weights are visible between the fields as two "preshapes" in the 2D field (also called *memory traces* which are subthreshold activity bumps), indicating, for two different intentions, which regions of CoS field they boost, if activated.

The intention peaks can be thought of as behavioral indices. A given behavior terminates once its associated CoS field goes above threshold. The CoS field gets input from the perceptual system (not shown), and is driven above threshold in the cases where the input stimuli match the preshape location.

After learning, one can see the effect of the weights, by referring back to Eq. 2. Based on how the intention field

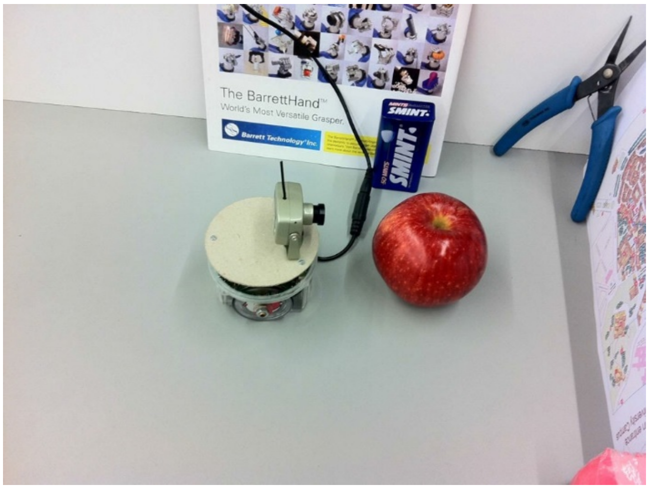

Figure 4: One of the E-Puck's intentions is satisfied by perception of the color red. If that intention were active, this would be a rewarding state for the robot. If the other intention were active, this would not be rewarding. When the reward signal is positive, all colors detected in the image are gradually associated with the CoS for that intention. It is essential for the robot to see different background colors. Of course, if one never sees a teacup apart from its saucer, one will never understand they are two separate objects.

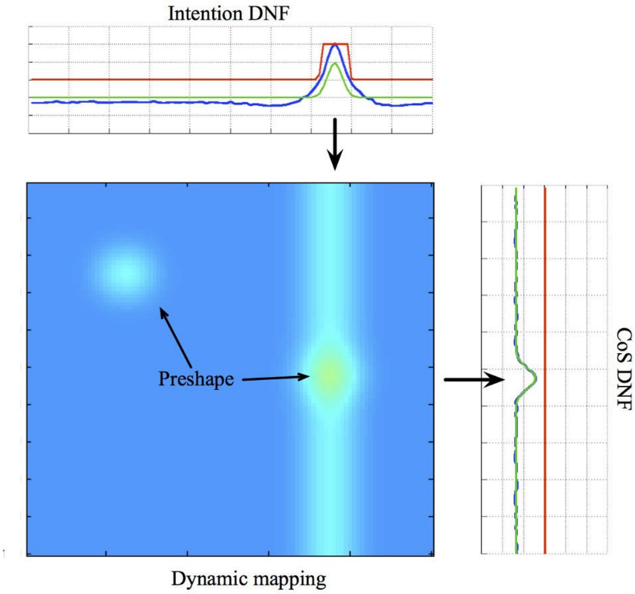

Figure 5: Example weighted mapping between one-dimensional intention and CoS dynamic neural fields.

output peak selects $x^{'}$ in the x dimension, and the corresponding y dimension (the CoS activity) is boosted according to $W(x^{'}, y)$.

In our simulations, function m in Eq. (2) which we call a "maturity" function, controls the transition between the learning phase to exploitation phase. m outputs a zero during a "guided learning" phase, in which intention has no effect on the CoS field. In this phase, external rewards from the teacher lead to peaks in the CoS field due to the boosts from these external rewards alone. In this phase, external reward is necessary for the weights to undergo learning. In the exploitation phase, $m$passes its input to its output so that the intention DNF biases the CoS field according to the learned weights.

The agent's learning of Wshould be mature enough, such that a CoS peak can result in the proper conditions without an external reward. The guided learning phase will be useful when the agent is "immature", either in the sense of being too young to have learned a proper W, or having learned an improper Wthrough some means, which now needs to be corrected. Alternatively, the weights could be used directly in both phases. In this case, the resting level of the CoS field should depend on the number of positive (learned) weights in the matrix W. In the beginning of learning, the strength of the summed weights is low and leads to a low resting level of the CoS DNF, which now cannot build activity peaks without the external reward (drive satisfaction). Later in the learning processes, the resting level of the CoS DNF is higher, so that the perceptual input and the weighted input from the intention field alone are enough for the activity peaks to be formed in the CoS DNF. Functionally, both these mechanisms are equivalent and here we choose a better controlled (but less autonomous) mechanism using the "maturity" function.

4 Implementation and Results

In order to illustrate the working of our learning mechanism, we present implementations on two robots – an E-Puck, and a Nao, with the latter tested in a simulated environment (using Webots [27]). The robots and their environments are shown in Fig. 3. Both robots receive visual input from their cameras through a visual perceptual DNF. This DNF spans over dimensions of color and location along the horizontal dimension of the image [15, 18] and builds activity peaks at positions that correspond to salient colored objects. Other feature dimensions have been used in other Dynamic Field Theory architectures [8], and could similarly be used with this mechanism as well.

The E-Puck was equipped with a new color camera (with higher frame-rate and resolution than the onboard camera), and was placed in a square enclosure, containing a red apple, a yellow block, and multi-colored distractor items and surrounding walls. The NAO humanoid robot was placed in front of a table with a pink block and a blue block, in front of a color-changing background wall.

Each robot switches between two elementary behaviors during learning. Activation of the respective intentions for the E-Puck was controlled by the teacher, through an interface. The NAO intentions were switched back and forth on a timer. Each EB intention did not initially have a defined Condition of Satisfaction, meaning the weight mapping was initially set to all zeros. These weights were learned over each experiment.

Whereas the E-Puck implementation did not use motor behavior, instead being controlled by the teacher, the NAO used a random 'babbling' motor behavior. More specifically, the E-Puck switched between various views, with different multi-color backgrounds, while the Nao switched between two focus points over a background surface that switched colors.

The learning process we described in Section 3 was utilized in both situations. The weight learning was gated by reward to associate the features (colors) that corresponded to the eventual satisfying condition. The E-Puck rewards were given by the teacher, while the NAO rewards were automated such that the reward was given as a constant signal for a short time after the intention and environment conditions matched.

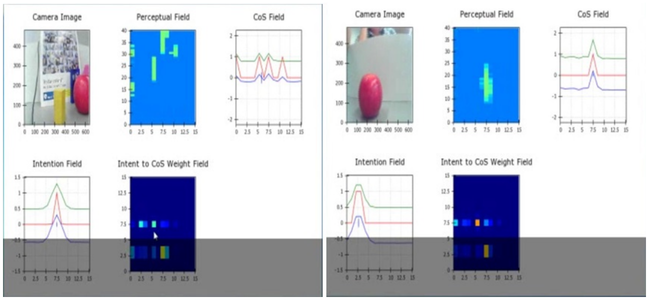

Figure 6: Snapshots of the E-Puck's dynamic fields during the learning. Left: The primary intention (drive) "thirst" is activated, which is satisfied by perception of the yellow color. When the rewarding signal is received, three colors are prevalent in the observed scene – yellow, red, and blue, and these all leave memory traces in the weights connecting the intention and the CoS DNFs. When learning continues and rewards are experienced in different scenes, the correct mapping activated, which is satisfied by the perception of the red color. Since only the red object is present in robot's view when the rewarding signal is received, a single peak is activated in the CoS field and only weights towards its location are strengthened.

4.1 Results of experiments on an E-Puck robot

The E-Puck was trained by a teacher in the real world, in real time. The robot had two intentions, each of which would be satisfied by a different color, but it did not know what these colors were initially. For the sake of discussion, we can label these drives 'hunger' and 'thirst'. The drives became active at different times: With the hunger drive active, reward was only obtained when a red object was in the image, seen in Fig. 4. When thirst was active, reward was obtained with a yellow object in the image. The actual reward was contingent on the teacher's input, through a training interface.

The robot was freely moved around the arena in a pseudo-random manner. The camera images provided input to a two-dimensional *perceptual field* [18], with one dimension as color hue (separated into 15 bins) and the other as the image columns. Along each column of the camera image, the hue of the pixels was summed to provide input to a certain location in the perceptual field. Activity peaks were formed in the perceptual field, detecting color objects along the horizontal dimension of the image. Positive activation in the perceptual field was projected onto the hue dimension and provided input to the CoS field. However, without either a reward signal, which uniformly boosts the CoS field, or a targeted boost (preshape) from the intention field, the CoS field cannot achieve super-threshold activation levels in order to generate an output peak.

The function of the teacher-provided reward signal was to provide this *boost* to the CoS field activation. Such a boost allows a peak to emerge in the output. As a result, the CoS field and intention field are simultaneously active, allowing the associative learning rule to adapt the weights between the active intention (corresponding to the active drive), and the CoS field.

Fig. 6 shows a snapshot of the system in action. The peak in the Intention Field reflects the currently active intention. In the Perceptual Field shown in the left screenshot, the colored objects lead to hue feature activations at yellow, red, and blue (white is not perceived as a color). Even though the color yellow in the center is the reason for a reward, all three colors become slowly associated with this intention. When the robot experiences the reward in many different contexts, the incorrect cues in the CoS weights are diminished over time. On the right is shown an uncluttered scene, for comparison.

One can see a video of the experiment at people.idsia.ch/~luciw/videos/epuckcos.wmv. After approximately 5 minutes of the experiment, with objects being moved around such that many contexts were experienced, the correct mappings were learned.

After the weight matrix is learned, the reward and the teacher became unnecessary to achieve satisfaction. The weights provided a sufficient boost to activate the CoS, and under the appropriate conditions, this boost would be selective for the perceptual conditions under which reward was achieved. The Condition of Satisfaction will work as needed in order to terminate its elementary behavior.

4.2 Results of experiments on a simulated NAO robot

The simulated NAO robot was tested in similar, but more automated, conditions than the EPuck. In particular, the robot "explored" the environment by looking left and right, with a timer causing the switch in head direction. A separate timer, which did not line up with the first, caused the switch between drive A and B. The system received a stream of visual inputs from the robot's camera. The camera images provided input to a two-

dimensional perceptual field (Hue × Column). Internal drives (as before, analogous to hunger and thirst), were structured such that a reward was only achievable by finding the object which is selectively rewarding for the currently active drive. When the NAO was motivated by Drive A, it could only achieve a reward by focusing on the pink object. When motivated by Drive B, it could only achieve a reward via the blue object.

Shots of the dynamic fields and weights, along with the environment, throughout the learning stages, are shown in Fig. 7. The reward signal provided a *boost* to the CoS field activation. This reward signal occurs when a drive is "satisfied" - drive A was satisfied by the perception of pink (Fig. 7(a)), but was not satisfied by the perception of blue (Fig. 7(b)). However, the background colors caused the weights as shown in part (b) to be as-yet non-selective. The weights are shown in the bottom two subfigures, and indicated by the blue line in the lower right subfigure. This was early in learning, however. Part (c) shows that after enough learning, the weights associated with drive A became selective for a single color (pink). A video of the learning is viewable at people.idsia.ch/~luciw/videos/naocosbefore.mov.

This basic exploration behavior along with the associative learning mechanism we described led to the learning of a weight matrix that appropriately encoded the Conditions of Satisfaction. Fig. 8 shows the robot after learning. Once the weight matrix was learned, the actual reward (and here, the teacher) became unnecessary, as the conditions of satisfaction were internalized. At this point, the weights provided a sufficient boost to activate the CoS, and this boost was selective for the perceptual conditions under which reward was achieved. (a): While drive A is active, the learned weights caused the large but sub-threshold peak in the perceptual field, which was further boosted by the perception of pink. The other, small, peak was due to the background color. (b): When drive B was active, a large but sub-threshold peak was caused by the weight matrix in the CoS field, for the color blue, which was pushed above the threshold by the perception of blue. A video of the NAO after learning can be viewed at people.idsia.ch/~luciw/videos/naocosafter.mov.

4.3 Implementation Details

With respect to implementation, neural fields were constructed using the Matlab package cosivina (https://bitbucket.org/sschneegans/cosivina/). 40 ms passed in Webots between each time step. The robot shifted its head every 40 time steps. The background wall changed to a randomized color every 15 time steps. The drive changed every 100 time steps. The time constant, τ, for the perceptual field was set to 7/3, while the intention and CoS fields halved that. Resting levels for the perceptual and intention field were set to −5, while the CoS field was set to −2.5. The sigmoidal slope value for all fields was set to 4. The image size was 120 × 160. The perceptual, intention and condition of satisfaction fields were set up for "find color" behaviors in the standard way [18], except the CoS weights were initially zeroed. The 2D perceptual field was composed of one dimension which was a mapping from hue to (0 20], with the second dimension as image column. The two hues of the two box objects were at hue values 13 and 17. The 1D CoS field was over the hue dimension, and took one of its inputs from the output of the perceptual field, projected over this hue dimension. The CoS field used neither local lateral interactions nor global inhibition. The 1D intention field was over values which had no external meaning. The intention field received a gaussian stimuli, with amplitude 5.5 and $\sigma = 1$, centered in a location dependent on the current drive, where drive A provided a stimuli centered at 5, and drive B provided a stimuli centered at 15. The intention field used local lateral interactions with a excitatory width parameter of 3, excitatory amplitude 15, inhibitory width 6.5 and inhibitory amplitude 15. No global inhibition was used. The intention field's output location set the row of the weight matrix, while the perceptual field output locations (multiple possible when multiple colors are in the image) set the columns. Over the points of intersection, learn-

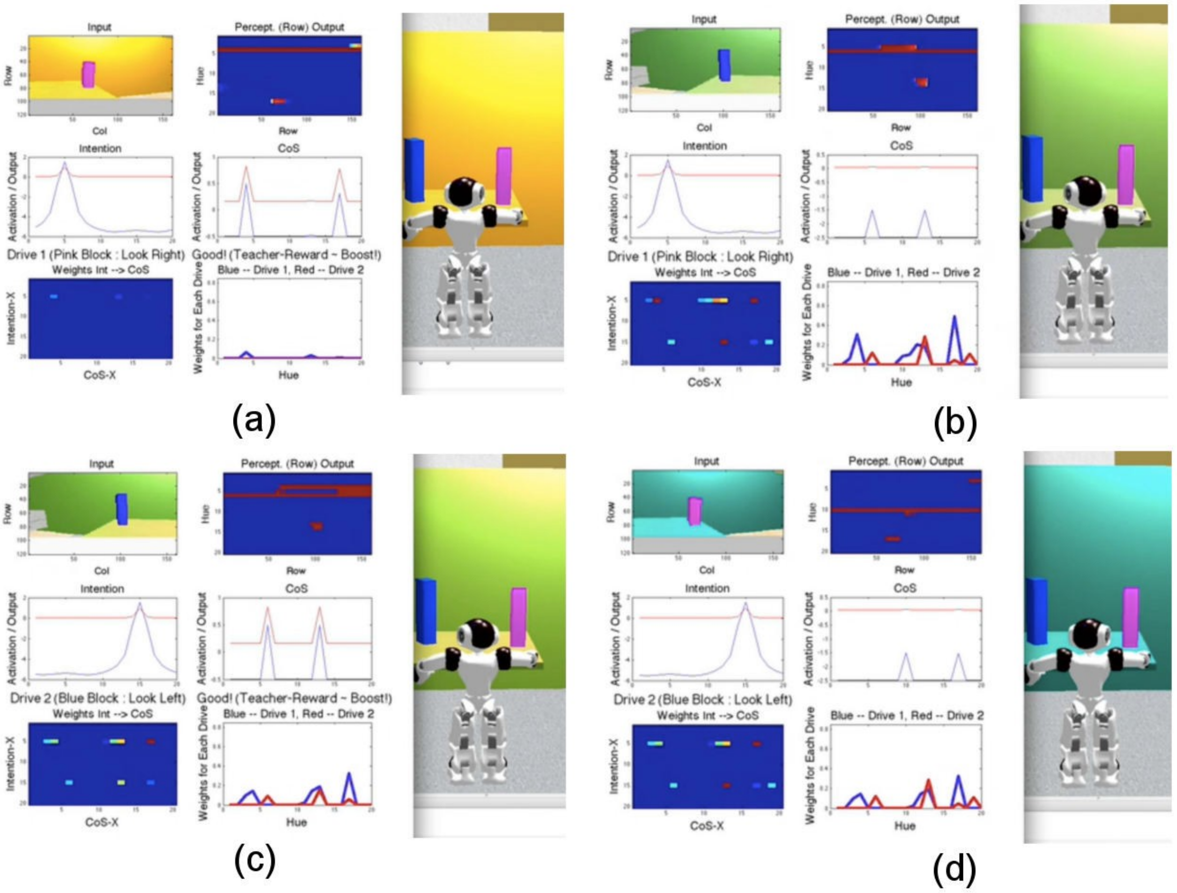

Figure 7: NAO during various stages of learning. With Drive A active, the NAO receives a reward when it finds a pink object as shown in (a), but not when it finds a blue object (b). When the reward is received in (a), weights from the intention to CoS field are boosted not only for the rewarding object color (pink), but incorrectly boosted for the background color as well. When Drive B is active, the NAO only receives a reward for finding a blue object, (c), but not for finding the pink object (d). As before, when a reward is received for finding the block that satiates the active drive, weights are not only boosted for the correct color, but for the incorrect background color as well. After learning over a large number of trials however, only the rewarding color weights remain, with the incorrect weights driven to 0 (shown in Fig. 8)

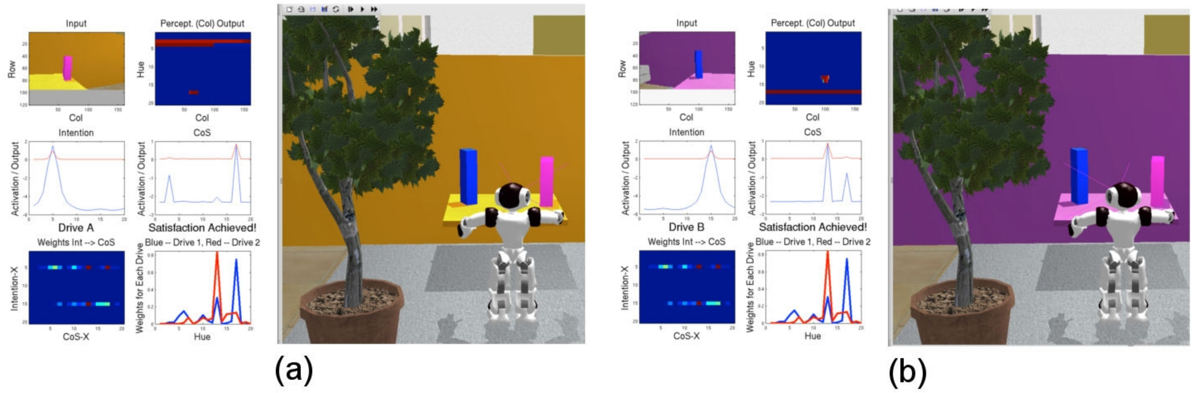

Figure 8: NAO after learning. After learning, the NAO only receives a boost in activation of the CoS field for the correctly rewarding color. When Drive A is active (shown in (a)), the CoS field is selectively excited for the pink object, while for Drive B (shown in (b)), it is selectively activated the blue object.

ing occurred, at a rate 0.01, when reward was available. Reward boosted the CoS resting level by 2, leading to super threshold activation for the colors in the image, whereby an association between current intention and CoS was reinforced. The values over the row of the weight matrix corresponding to current intention were input to the CoS field as a secondary input, along with the perceptual field output, after maturity. During this post-learning phase, the weight matrix projection to CoS was multiplied by 2.

5 Discussion

5.1 Relationship with Reward Prediction

When an elementary behavior is rewarded upon completion, its CoS field is, in a sense, a *reward predictor*, due to the short delay between the agent sensing the correct conditions, i.e., the emergence of a CoS peak, and actually perceiving the reward. Learning to anticipate the outcome of an action has been extensively discussed and there are many existing biologically-plausible reward prediction learning mechanisms that handle the case of predicting immediate reward [22, 23]. Other reward prediction methods go beyond one-step prediction and are not directly related to the animal learning literature [26]. In such reinforcement learning approaches, the state or state-action value function associated with a policy is a reward predictor with a discounted infinite horizon. Schmidhuber, for example, considered reinforcement as another type of input [20], and the non-discounted prediction and acquisition of this reward was managed by a fully recurrent dynamic control network.

5.2 Relationship with Classical Conditioning

Similar learning processes to those undergone in the experiments have been studied in animal behavioural experiments, in particular using different conditioning paradigms [4]. For instance, in instrumental conditioning, the animal learns the association between the desired outcome and the selected action [14]. An explicit representation of expected outcomes of actions is emphasized in experiments on Differential Outcome learning.

In the model presented here, the CoS learning process is related to such conditioning experiments, in which animals learn to associate satisfaction of a certain basic drive – hunger or thirst – with the outcome of a particular action. By doing this, we try to answer the question: what are the origins of elementary behaviors? We consider, in general, one of the origins to be *endogenous drives*. The drives here follow the definition provided by Woodworth [30], who explicitly distinguished the notion of 'drive' from 'mechanism'. Whereas 'mechanisms' refer to *how* an agent can achieve a goal, 'drives' refer to *why* one might want to achieve a goal in the first place. As prototypical examples of bodily drives, Woodworth suggested hunger and thirst, each of which serve as internal forces for motivating various sorts of behaviors [10]. The method presented here enables an agent, motivated by a set of such drives, to learn to recognize the perceptual conditions associated with desirable outcomes.

We have demonstrated how drive satisfaction may lead to development of an anticipatory representation of the outcome of an action. In neural-dynamic terms, the coupling between intention and condition of satisfaction of an elementary behavior is learned. After such learning, the agent may detect a successful accomplishment of an action without the need for an externally provided drive-satisfaction signal. This anticipatory representation of the final state of the action may be used to drive activation of the next item in a behavioural sequence [13].

6 Conclusions

In this work, we show a Dynamic Neural Field-based architecture that allows the learning of a coupling between the intention of an action and its condition of satisfaction. This coupling amounts to an anticipation of the outcome of the action and is learned based on rewarding signals, received when an internal drive such as hunger or thirst) is satisfied. After learning, the perception of the CoS is enough for the agent to perceive the action as finished, external (to the nervous system) reward is not needed any more. The method enables both a realworld, E-Puck robot, and a simulated NAO humanoid robot to learn the conditions of satisfaction for different behaviors, in their respective environments.

The Dynamic Neural Fields, used to implement intentions and CoS of the agent's behaviours are continuous activation functions, defined over the relevant feature spaces. Thus, the location of the activation peak in this field is determined by the current sensory input, which drives these fields. Moreover, the peaks have finite width and consequently, the learned coupling between the intention and the CoS DNFs (1) reflects the actual sensory state, experienced by the agent during learning and (2) generalises to neighbouring locations in the feature dimension. If during learning the activity peaks were experienced over several neighbouring locations in the CoS field, the weight matrix will reflect the experienced distribution of peaks, although with less “certainty" (strength of respective weights).

This work is the first step towards learning elementary behaviours, which structure the behavioural repertory of an embodied agent and control its behaviour. The model demonstrates how the association between the intention and the anticipated condition of satisfaction may be learned based on sensory input and unspecific rewarding signal in a behaving agent.

Acknowledgement: The authors gratefully acknowledge the financial support of the European Union Seventh Framework Programme FP7-ICT-2009-6, under Grant Agreement no. 270247 “NeuralDynamics”, and of the DFG SPP Grant “Autonomous Learning”, within Priority Program 1567.

References

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

[2] E Bicho, P Mallet, and G Schöner. Target representation on an autonomous vehicle with low-level sensors. *The International Journal of Robotics Research*, 19:424–447, 2000.

[3] Daniele Caligiore, Anna M Borghi, Domenico Parisi, and Gianluca Baldassarre. Tropicals: a computational embodied neuroscience model of compatibility effects. *Psychological Review*, 117(4):1188, 2010.

[4] Anthony Dickinson and Bernard Balleine. Motivational control of goal-directed action. *Animal Learning & Behavior*, 22(1):1–18, 1994.

[5] B Duran and Y Sandamirskaya. Neural dynamics of hierarchically organized sequences: a robotic implementation. In *Proceedings of 2012 IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2012.

[6] Boris Duran, Yulia Sandamirskaya, and Gregor Schöner. A dynamic field architecture for the generation of hierarchically organized sequences. In Alessandro E.P. Villa, Wlodzislaw Duch, Peter Erdi, Francesco Masulli, and Günther Palm, editors, *Artificial Neural Networks and Machine Learning – ICANN 2012*, volume 7552 of *Lecture Notes in Computer Science*, pages 25–32. Springer Berlin Heidelberg, 2012.

[7] Wolfram Erlhagen and Estela Bicho. The dynamic neural field approach to cognitive robotics. *Journal of Neural Engineering*, 3(3):R36–R54, 2006.

[8] C Faubel and G Schöner. Fast learning to recognize objects: Dynamic fields in label-feature space. In *Proceedings of the fifth International Conference on Development and Learning ICDL 2006*, 2006.

[9] S Grossberg. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Networks*, 1:17–61, 1988.

[10] C.L. Hull. *Principles of behavior: an introduction to behavior theory*. Century psychology series. D. Appleton-Century Company, incorporated, 1943.

[11] 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*, 26:227–251, 2008.

[12] S. Kazerounian, M. Luciw, M. Richter, and Y. Sandamirskaya. Autonomous reinforcement of behavioral sequences in neural dynamics. In *International Joint Conference on Neural Networks (IJCNN)*, 2013.

[13] Giovanni Pezzulo, Martin V Butz, and Cristiano Castelfranchi. The anticipatory approach: definitions and taxonomies. In *The Challenge of Anticipation*, pages 23–43. Springer, 2008.

[14] R A Rescorla and R L Solomon. Two-process learning theory: Relationships between pavlovian conditioning and instrumental learning. *Psychological Review*, 74(3):152–182, 1967.

[15] Mathis Richter, Yulia Sandamirskaya, and Gregor Schöner. A robotic architecture for action selection and behavioral organization inspired by human cognition. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2457–2464, 2012.

[16] Y. Sandamirskaya, M. Richter, and G. Schöner. A neuraldynamic architecture for behavioral organization of an embodied agent. In *IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL EPIROB 2011)*, 2011.

[17] Yulia Sandamirskaya. Dynamic neural fields as a step towards cognitive neuromorphic architectures. *Frontiers in Neuroscience*, 7:276, 2013.

[18] Yulia Sandamirskaya and Gregor Schöner. An embodied account of serial order: How instabilities drive sequence generation. *Neural Networks*, 23(10):1164–1179, December 2010.

[19] Yulia Sandamirskaya, Stephan K.U. Zibner, Sebastian Schneegans, and Gregor Schöner. Using dynamic field theory to extend the embodiment stance toward higher cognition. *New Ideas in Psychology*, 2013.

[20] J Schmidhuber. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Institut für Informatik, Technische Universität München. Technical Report FKI-126-90*, 1990.

[21] G Schöner. Dynamical systems approaches to cognition. In Ron Sun, editor, *Cambridge Handbook of Computational Cognitive Modeling*, pages 101–126, Cambridge, UK, 2008. Cambridge University Press.

[22] Wolfram Schultz. Reward signaling by dopamine neurons. *The Neuroscientist*, 7(4):293–302, 2001.

[23] Wolfram Schultz. Review dopamine signals for reward value and risk: basic and recent data. *Behav. Brain Funct*, 6:24, 2010.

[24] John R Searle. *Intentionality — An essay in the philosophy of mind*. Cambridge University Press, 1983.

[25] J P Spencer and G Schöner. Bridging the representational gap in the dynamical systems approach to development. *Developmental Science*, 6:392–412, 2003.

[26] R.S. Sutton and A.G. Barto. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press, 1998.

[27] Webots. http://www.cyberbotics.com. Commercial Mobile Robot Simulation Software.

[28] Juyang Weng. Developmental robotics: Theory and experiments. *International Journal of Humanoid Robotics*, 1(02):199–236, 2004.

[29] H R Wilson and J D Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik*, 13:55–80, 1973.

[30] R.S. Woodworth. *Dynamic psychology, by Robert Sessions Woodworth*. Columbia University Press, 1918.