A Dynamic Neural Field Approach to Natural and Efficient Human-Robot Collaboration

Wolfram Erlhagen and Estela Bicho

chapter 13 in Neural Fields: Theory and Applications, pp 341-365, Edts Stephen Coombes, Peter beim Graben, Roland Potthast, and James J. Wright, Springer Berlin Heidelberg 2014 , 31st May 2014, Print ISBN 978-3-642-54592-4, Online

Abstract A major challenge in modern robotics is the design of autonomous robots that are able to cooperate with people in their daily tasks in a human-like way. We address the challenge of natural human-robot interactions by using the theoretical framework of dynamic neural fields (DNFs) to develop processing architectures that are based on neuro-cognitive mechanisms supporting human joint action. By explaining the emergence of self-stabilized activity in neuronal populations, dynamic field theory provides a systematic way to endow a robot with crucial cognitive functions such as working memory, prediction and decision making . The DNF architecture for joint action is organized as a large scale network of reciprocally connected neuronal populations that encode in their firing patterns specific motor behaviors, action goals, contextual cues and shared task knowledge. Ultimately, it implements a context-dependent mapping from observed actions of the human onto adequate complementary behaviors that takes into account the inferred goal of the co-actor. We present results of flexible and fluent human-robot cooperation in a task in which the team has to assemble a toy object from its components.

1 Introduction

Recent advances in robotics technology make the design of socially interactive robots that work closely with ordinary people in their day-to-day work a realistic goal (Fong et al., 2003). Research in such human-centered robotics requires to address a wealth of new interdisciplinary topics from cognitive psychology, artificial

Wolfram Erlhagen
Department of Mathematics and Applications, Center for Mathematics, University of Minho, Portugal e-mail: wolfram.erlhagen@math.uminho.pt

Estela Bicho
Department of Industrial Electronics, Centre Algoritmi, University of Minho, Portugal e-mail: estela.bicho@dei.uminho.pt

intelligence and neuroscience that go well beyond traditional mathematical issues of robotics research for industrial applications (Schaal, 2007). As fundamentally social beings, we are experts in joint activity in order to realize a common goal. We therefore have high expectancies about an engaging and pleasant interaction with another agent. Humans achieve their remarkable fluent organization of joint activity in routine tasks, such as preparing the dinner table, by continuously monitoring the partner's actions, and predicting them effortlessly in terms of their outcomes (Sebanz et al., 2006). Based on this prediction, an adequate complementary action can be timely selected among all potential behaviors that the task currently affords. To ensure user acceptance, a socially interactive robot that is supposed to substitute a human in a cooperative task should equally contribute to the coordination and synchronization of behaviors among the co-actors. It is thus crucial to endow the robot with high-level cognitive functions such as action understanding, decision making and memory.

Given the large variety of disciplines involved in the emerging field of human-friendly robotics, it is perhaps not surprising that different design approaches toward more natural human-robot interaction have been proposed. Conceptually, they may be broadly classified in top-down, symbolic views on human-like (social) intelligence and more bottom-up, neurodynamics and embodied notions (Kozma, 2008). The predominant top-down approach is inspired by traditional artificial intelligence (AI) models that address the complex problem of selecting an adequate complementary behavior as a sequence of logical operations performed on discrete symbols. The robotics implementations are thus based on formal logic and formal linguistic systems (Levesque and Lakemeyer, 2008). Good examples are architectures inspired by the theoretical framework of joint intention theory (Cohen and Levesque, 1990; Alami et al., 2005; Hoffman and Breazeal, 2007). This framework provides a rigorous logical treatment of how sub-plans of individual agents committed to a common task can be meshed into joint activity. A defining feature of the symbolic approach is that information processing is set up in stages from perception to cognition to action. A perceptual subsystem first converts sensory information about external events into inner symbols to represent the state of the world. Next, this information is used along with representations of current goals, memories of past events and beliefs about the partner's intention to decide about the course of action. On this planning level, actions are formulated as logical operators with preconditions and effects that change the world in a discrete fashion and instantaneously. The abstract plan is then transformed into motor representations of the robotics system that are finally used to generate arm and hand trajectories in order to realize the plan.

The symbolic, disembodied view on how to decide what to do has provided many impressive examples of intelligent behaviors in artificial agents (for review see Vernon et al. (2007)). However, it is now widely recognized by the robotics and cognitive science communities that the symbolic framework based on serial stages of processing has notoriously problems to cope with real-time interactions in dynamic environments (Haazebroek et al., 2011; Levesque and Lakemeyer, 2008; Kozma, 2008). In human-robot interaction tasks, the robot has to reason about a world that may change at any instance of time due to actions taken by the user. Even if we consider that the processing in the perceptual and decision modules would allow to continuously update the robot's plan in accordance with the user's intention, the extra processing step needed to embody the abstract action plan in the autonomous robot would challenge the fluent and seemingly effortless coordination of decisions and actions that characterize human joint action in familiar tasks.

In order to advance toward a more online view of high-level social cognition, our group at the University of Minho has developed and tested over the last couple of years a neurodynamics approach based on the theoretical framework of Dynamic Neural Fields (DNF)(Erlhagen and Bicho, 2006). The DNF model for natural human-robot interaction that we present in this chapter implements known neuro-cognitive processing mechanisms supporting dynamic social interactions in humans and other primates (Sebanz et al., 2006). Converging lines of experimental evidence in behavioral and neuro-cognitive studies suggest that the interaction between sensory, cognitive and motor processes in the brain is much more interactive and integrated as previously thought. For instance, neural correlates of decision making seem to be inconsistent with the notion that a central decision maker completes its operation before activating the motor structures to perform the action plan (Gold and Shadlen, 2007). Instead, the process of action selection may be best understood as a winner-takes-all competition between multiple neuronal population representations of motor behaviors that the environment currently affords (Cisek, 2007). The advantage of such a dynamic competition process for flexible behavior is obvious. Since the flow of sensory information is continuously used to partially specific several potential actions, the system is prepared to quickly adjust to a changing world. Different neural pathways carrying different sources of information demonstrate the tight coupling between visual and motor systems (for review see Rizzolatti and Luppino (2001)). For instance, according to the concept of object affordances (Gibbson, 1979), the perception of a graspable object immediately activates to some extent the neuronal representations of potential motor interactions with that object. The final decision to execute a certain action, represented by a sufficiently activated subpopulation, may depend on additional contextual cues and the current behavioral goal. Very important for social interactions, an impressive body of experimental evidence from behavioral and neurophysiological studies investigating action and perception in a social context shows that when we observe other's actions corresponding motor representations in our motor system become activated (for a recent review see Rizzolatti and Sinigaglia (2010)). In a cooperative joint action context like transferring an object to a partner, this automatic action resonance mechanism has been interpreted as evidence that the likelihood of performing a complementary motor program is increased, that is, the 'receiver' immediately prepares a complementary grasping behavior that ensures a safe and robust object transfer (Newman-Norlund et al., 2007). For more complex joint action settings for which the mapping from observed actions onto adequate complementary behaviors is not as clear, the observer has first to predict the partner's ongoing action in terms of the future effects in the environment. The action resonance mechanism is believed to support also the high-level cognitive functionality of action understanding and goal inference (Rizzolatti and Sinigaglia, 2010). The key idea here is that the observer internally simulates the outcome of perceived actions using his/her own motor representations that have become associated with representations of action goals during learning and practice. The notion that motor representations are crucially involved in a higher-cognitive function like generating expectations about the future is clearly inconsistent with serial information processing theories of cognitive behavior.

The DNF model of cooperative joint action is organized as a large scale network of reciprocally connected neuronal populations that encode in their firing patterns specific motor behaviors, action goals, contextual cues and shared task knowledge (Bicho et al., 2011a,b). Although some level of functional modularity exists in the network, it is important to notice that the formation and maintenance of a behavioral decision is not represented in the discharge pattern of "motor" neurons alone, but is distributed among all currently active populations in the network.
The activity in each local population evolves continuously in time under the influence of external input from connected neuronal pools or the sensory system and recurrent excitatory and inhibitory interactions within the population. Central for the design of cognitive agents, the recurrent interactions support the existence of self-sustained bumps of activation. Persistent population activity allows us for instance to implement a working memory function in the robot to cope with temporally missing sensory information, or to simulate future environmental inputs that may inform the current decision process about a goal-directed behavior (Erlhagen and Bicho, 2006).
As a specific mathematical formulation of a DNF, we adopt Amari's model for pattern formation in neural populations since it allows analytical treatment (Amari, 1977). This is an important advantage when trying to design a complex robot control architecture for real-world experiments.

The chapter is organized as follows: first, we give an overview about the neuro-cognitive foundations of the DNF model and describe its mathematical implementation. We then illustrate the coordination of actions and decisions between human user and robot organized by the network dynamics in a joint action task in which the two teammates have to jointly assemble a toy object from its components.

2 Dynamic Neural Field Model of Joint Action

As a working definition, joint action can be regarded as any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment. Crucial building blocks for successful joint action coordination are the capacities to recognize actions performed by others, and to integrate predicted effects of own and others' behaviors in the action selection process (Sebanz et al., 2006). What are the neural bases of efficient social interactions? The discovery of the so-called mirror neuron system first in monkey and later in human gives strong support for the hypothesis that observing actions performed by another individual elicit a motor activation in the brain of the observer similar to that which occurs when the observer plans his/her own goal-directed action (for a recent review see Rizzolatti and Sinigaglia (2010). This automatic action resonance mechanism has given rise to the hypothesis that covert motor simulations support action understanding in a social context without the costs that are associated with conscious mental processes or explicit communication.

Mirror neurons in premotor cortex of monkeys (area F5) become active both when the monkey performs a specific motor act like grasping an object and when it observes another individual making a similar action. Importantly, for most mirror neurons the congruency between the observed and the executed motor act is relatively broad. This suggests that their discharge is not related to the fine details of the movements but codes the goal of the observed or executed motor act. Object manipulation tasks typically involve a series of action phases like reaching, grasping, lifting, holding and placing that are bounded by specific sensory events defining subgoals of the task (Flanagan et al., 2006). Distinct populations of mirror neurons are assumed to represent the functional goals of these successive action phases. Mirror neurons have been also described in areas PFG and PF of the inferior parietal lobe (IPL). These areas are anatomically connected with premotor area F5 and with higher visual areas in the superior temporal sulcus (STS). STS neurons discharge during hand-object interactions similar to those encoded by F5 neurons. The difference seems to be that STS neurons do not discharge during overt movements. STS neurons thus might provide mirror neurons with a visual description of goal-directed motor acts.

The hypothesis that the discharge of neuronal populations in the STS-PFG/PF-F5 circuit play a key role in action understanding and goal inference has obtained strong support from a series of neurophysiological experiments. It has been shown for instance that grasping mirror neurons are activated also when the critical part of the observed action, the hand-object interaction, is hidden behind a screen and can thus only be inferred from additional contextual information (e.g., the presence of a graspable object behind the occluding surface (Umiltà et al., 2001)). In a recent study, Fogassi and colleagues (2005) reported that IPL mirror neurons, in addition to recognize the goal of an observed motor act, discriminate identical grasping behaviors according to the final goal of the action sequence in which the motor act is embedded (e.g., grasping for eating versus grasping for placing in a container). They further argued that because the discriminated motor act is part of a specific chain of motor primitives associated with a specific goal representation most likely in prefrontal cortex (PFC), the monkey could predict at the time of the grasping the ultimate goal of the observed action and, thus read the motor intention of the acting individual. Of course, the discrimination of the grasping behavior is only possible because of an additional contextual cue (e.g., the presence of a container in the scene). This sugge~~st~~ that the simulation process in IPL mirror neurons is not exclusively shaped by input from STS but also depends on input from goal and object representations.

Figure 1 sketches the multi-layered dynamic field model of joint action consisting of various neural populations that are associated through hand-coded synaptic links (not all are shown to avoid crowding). As a central part, it integrates a previous DNF model of action understanding and goal-directed imitation inspired by the mirror system(Erlhagen et al., 2006). Ultimately, the distributed network implements a flexible mapping between observed and executed actions that takes into account the inferred goal of the co-actor, contextual cues and shared task knowledge.

——— Insert Figure 1 around here ———

An observed hand movement that is recognized by the vision system as a particular movement primitive (e.g. a whole hand-grasping-from above) is represented by suprathreshold activity of a specific neuronal population in the action observation layer (AOL). Input from AOL to corresponding populations in the action simulation layer (ASL) may activate together with input from the object memory layer (OML) and the common sub-goals layer (CSGL) specific chains of movement primitives that are linked to neuronal representation of the ultimate action goal in the intention layer (IL)(Erlhagen et al., 2007). Suprathreshold population activity in IL will drive one or more associated populations in the action execution layer (AEL) that represent possible complementary motor behaviors. Similar to ASL, the motor behaviors are organized in chains of motor primitives like reaching-grasping-placing. There are different ways how to represent the temporal order and the timing of motor sequences in the dynamic field framework (Ferreira et al., 2011; Sandamirskaya and Schöner, 2010). To simplify the present robotics experiments with its emphasis on competitive action selection, we have not modeled these chains as a sequential activation of individual neural populations, but represent the entire motor behavior by a single pool of neurons. The final decision in AEL depends not only on the input from IL but also on input from OML and CSGL. OML contains neuronal population representations of the various objects in the scene. It is organized in two layers that discriminate whether a specific object is within the user's or within the robot's reachable space. Input from OML automatically pre-activates neural representations of associated motor behaviors in AEL. Specifically for the joint assembly task, possible object-directed behaviors include the transfer of the object to the co-actor or a direct placement of the object as part of the assembly work. In addition, communicative gestures like for instance pointing to the specific component may be used in joint activity to attract the co-actor's attention (Bicho et al., 2010). Efficient task performance requires to carry out the steps in the task in the correct order, without repeating an action or omitting early actions in the sequence. This behavioral planning heavily depends on the predicted consequences of intended actions (i.e. a change in the state of the target object (Tanji et al., 2007)). The common subgoals layer CSGL contains neuronal representation of desired end results of individual assembly steps that can be realized by associated motor representations in AEL and that are recognized by the vision system. Neurophysiological evidence suggests that in sequential tasks, distinct subpopulations in PFC represent already achieved subgoals and subgoals that have still to be accomplished (Genovesio et al., 2006). In line with this finding, CSGL contains two connected DNF layers with population representations of past and future events. Input from the vision system about the achievement of a specific subgoal activates the corresponding population in the past layer, which in turn inhibits the corresponding goal representation and simultaneously excites one or more populations in the future layer. They represent in their activity patterns predicted end result of subsequent assembly steps that the current state of the assembly work allows. Important for the fluency of the team behavior, the updating of subgoals in CGSL may not only be triggered by direct input from the vision system but also by input from IL representing the inferred motor intention of the co-actor. This allows the observer to prepare future actions in response to anticipated rather than observed action outcomes (Bicho et al., 2011a,b).

3 Model Details

In their seminal work, Wilson and Cowan (1973) and Amari (1977) introduced dynamic neural fields as rate models of cortical population dynamics that abstract from the biophysical details of neural firing. The architecture of this model class reflects the hypothesis that strong excitatory and inhibitory interactions within local populations that receive synaptic input from multiple connected neuronal pools form a basic mechanism of cortical information processing. As shown in numerous simulation studies, dynamic neural field models are powerful enough to reproduce neural population dynamics observed in neurophysiological experiments (e.g., Erlhagen et al. (1999)), and to understand the basic mechanisms underling a large variety of experimental findings on the perceptual and behavioral level (for review see Schöner (2008)).

For the design of the robot control architecture for natural human-robot interactions, we adopt the model of a single layer of a homogeneous neural network consisting of excitatory and inhibitory neurons proposed by Amari (1977). This model allows for a rigorous analysis of the existence and stability of characteristics solutions such as local excitations or "bumps". In the following, we give a brief overview about the techniques developed by Amari, and explain the adaptations we have made to cope with the specific needs of the robotics implementations.
The dynamics of each population in the distributed network shown in Fig. 1 is governed by the equation:

$$\tau \frac{\delta u(x,t)}{\delta t} = -u(x,t) + S(x,t) + \int_{\infty}^{\infty} w(x-x') f(u(x',t)) dx' - h \tag{1}$$

where $u(x,t)$ is the average activity of neuron $x \in (-\infty, +\infty)$ at time t and parameter $\tau > 0$ defines the time scale of the field dynamics. The globally inhibitory input

$h > 0$ determines the resting state to which the activity of neuron x relaxes without external input $S(x,t) \geq 0$. The integral term in Eq. 1 describes the interactions within the populations which are chosen of lateral-inhibition type:

$$w(x) = A\exp(-x^2/2\sigma^2) - w_{inhib} \tag{2}$$

where $A > 0$ and $\sigma > 0$ describe the amplitude and the standard deviation of a Gaussian, respectively. For simplicity, the long-range inhibitory interactions are assumed to be constant, $w_{inhib} > 0$, implementing a competition between subpopulations that are sufficiently separated in space. Note that distinct neural populations encoding entire temporal behaviors like grasping, holding or placing seem to be spatially segregated in the mirror neuron areas (Rizzolatti and Luppino, 2001). Interpreting the metric of neural interactions in anatomical space like in Amari's original model is thus possible. However, the metric distance might be also defined in an abstract psychological space (Shepard, 1997). In this case, functionally distinct behaviors associated with specific goals would be represented by spatially separate, competing pools of neurons whereas similar motor behaviors associated with the same goal (e.g., grasping with different grip types) would be represented by partially overlapping populations.
Amari assumes for his analysis of pattern formation that the output function $f(u)$,which gives the firing rate of a neuron with input u, is the Heaviside step function, i.e., $f(u) = 0$ for $u \leq 0$ and $f(u) = 1$ otherwise. To model a more gradually increasing impact of the recurrent interactions on the population dynamics we apply a smooth and differentiable output function of sigmoid shape with slope β and threshold u_0:

$$f(u) = \frac{1}{1 + exp\left(-\beta(u - u_0)\right)}. \tag{3}$$

It has been shown by Kishimoto and Amari (1979) that many of the results concerning the existence and stability of localized activity patterns obtained with a step output function take over to the more general case of the sigmoid.

The model parameters are chosen to guarantee that the population dynamics is bi-stable, that is, the attractor state of a stable "bump" coexists with a stable homogeneous resting state. A sufficiently strong transient input $S(x,t)$ may drive the neural population beyond threshold, $f(u) > u_0$. The resting state loses stability and a localized activation pattern evolves. In the various layers of the network model, these bumps represent memorized information about object location, the inferred action goal of the co-actor or a decision for a specific complementary behavior. Weaker external input signals from connected populations lead to a subthreshold activation pattern for which the contribution of recurrent interactions is negligible. It is important to note, however, that this preshaping by weak input may nevertheless influence the robot's behavior. Since the level of pre-activation affects the rate at which a suprathreshold activation pattern rises (Erlhagen and Schöner, 2002), a pre-activated population has a computational advantage over a population at resting

level and thus has a higher probability to influence the decision process in AEL.
For the case of a step output function, the conditions for the existence and stability of a single bump in the presence of a stationary external input $S(x)$ can be easily derived following Amari's approach. Let $R(u) = \{x | u(x) > 0\}$ be the excited region of the field. A localized pattern of length $a = x_2 - x_1$ is then defined by the finite interval $R(u) = (x_1, x_2)$. Since at equilibrium $\frac{\delta u}{\delta t} = 0$ in Eq. 1, the equilibrium solution $\tilde{u}(x)$ satisfies

$$\tilde{u}(x) = \int_{R(\tilde{u})} w(x - x') f(\tilde{u}(x', t)) dx' - h + S(x) \tag{4}$$

By defining the function

$$W(x) = \int_0^x w(x') dx' \tag{5}$$

we have for the local excitation with $R(u) = (x_1, x_2)$

$$\tilde{u}(x) = W(x - x_1) - W(x - x_2) + S(x) - h \tag{6}$$

Since $W(0) = 0$ and $W(x) = -W(-x)$ the equilibrium local excitation with $\tilde{u}(x_1) = 0 = \tilde{u}(x_2)$ satisfies:

$$S(x_i) = -W(a) + h, \;\; i = 1, 2 \tag{7}$$

For the robotics experiments we are specifically interested in the existence of localized excitation in response to symmetric, bell-shaped input. In this case, the length a of the bump satisfies

$$S(x_0 + a/2) = h - W(a) \tag{8}$$

where x_0 denotes the position of the maximum $S(x)$. If $h > 0$ is chosen such that

$$W_m = \max_{x>0} W(x) > h \tag{9}$$

holds, there exist two solutions $\hat{a}$ and a, with $\hat{a} < a$, of Eq. 7. Amari reduces the neural field equation to an ordinary differential equation with respect to the boundaries of the excited region and uses a perturbation approach to show that only the larger excitation pattern is stable (for details see Amari (1977)).
For the robotics implementations we assume that the time dependent input from a connected population u_j to a target population u_i has a separable form $S_i(x,t) = S(x)g_j(t)$ where S(x) is modeled as a Gaussian function and $g_j(t) = 1$ if $f(u_j) > u_0$ and $g_j(t) = 0$ otherwise. In other words, a stationary input is applied during the period of suprathreshold activity in u_j. Numerical studies show that the evolving localized activation in u_j could have been directly used as input pattern as well. However, assuming a constant input shape allows us to closely follow Amari's analysis. The total input from all connected populations and external sources (e.g., vision system, also modeled as Gaussian signal) to u_i is then given by

$$S_\mathrm{i}(x,t) = k \sum_\mathrm{j} g_\mathrm{j}(t) A_\mathrm{j} \exp(-(x - x_\mathrm{i})^2 / 2\sigma^2) \tag{10}$$

where $k > 0$ is a scale factor to guarantee that the total external input remains small compared with the recurrent interactions within the local population.

To model different cognitive functions like working memory or decision making in the various layers of the model, we specifically adapt the basic field equation given by Eq. 1 accordingly. To implement in OML, AOL and CSGL a working memory function, it is important that a bump remains after cessation of the transient stimulus that has initially driven its evolution. The condition $W_m > h > 0$ guarantees the existence of a stable bump for $S(x) = 0$ which, however, has a slightly smaller width compared to the bump in the presence of input. We call this solution self-sustained to distinguish it from a suprathreshold activity pattern that becomes self-stabilized only because of the presence of external input. In this case, equation $S(x_0 + a/2) = h - W(a)$ has a solution which represents a stable localized activation but $h > W_m$ holds, that is, the field dynamics is in the mono-stable regime and suprathreshold activity will decay to rest state without external support.
To represent and memorize simultaneously multiple items, a multi-bump solution is required. An interaction kernel with long-range, constant inhibition (Eq. 2) may sustain multiple localized activity patterns without external inputs with additional stabilization mechanisms (Trappenbeg and Standage, 2005; Erlhagen and Bicho, 2006). For simplicity, we have used for the current robotics experiments kernels with limited spatial range to exclude mutual competition between multiple memories. An alternative solution that we are currently exploring for the robotics work is to use coupling functions with multiple zero-crossings, modeling excitatory interactions also at larger distances (Laing et al., 2002; Ferreira et al., 2011).
The memory is continuously updated in accordance with input from the vision system indicating a change in the external world (e.g., a new location of a specific object). To implement the “forgetting” process, we use a simple first-order dynamics with an appropriate time scale for the (local) adaptation of the inhibitory input h to destabilize an existing bump (Bicho et al., 2000):

$$\frac{dh}{dt} = -r_{h,min} c_h (h - h_{min}) - r_{h,max}(1 - c_h)(h - h_{max}) \tag{11}$$

where $|h_{max}| < W_m$ and $|h_{min}| > W_m$ are the two limit values for h that define the bi-stable and the mono-stable regime, respectively. The rate of change for destabilizing a memory function in case of an existing bump ($c_h = 1$) or restoring in the absence of a bump ($c_h = 0$) is given by the parameters $r_{h,min} > 0$ and $r_{h,max} > 0$.
To meet the real-time constraints of action selection and goal inference in a continuously changing environment, we apply in ASL, AEL and layer CSGL representing future subtaks a field dynamics with self-stabilized rather than self-sustained activation patterns. A decision to select a certain motor behavior that takes into account the most likely goal of the co-actor’s current action, is temporally stabilized by sufficient strong support of external and internal evidence, but will automatically lose stability if this evidence changes in favor of a competing behavior.

4 Setup of Human-Robot Experiments

To test the dynamic neural field model of joint action in human-robot experiments, we have adopted a joint assembly paradigm in which the team has to construct a toy 'vehicle' from components that are initially distributed on a table (Fig. 2).

——— Insert Figure 2 around here ———

The toy object consists of a round platform with an axle on which two wheels have to be attached and each fixed with a nut. Subsequently, four columns that differ in their color have to be plugged into corresponding holes in the platform. The placing of another round object on top of the columns finishes the task. The components were designed to limit the workload for the vision and the motor system of the robot. It is assumed that each teammate is responsible to assemble one side of the toy. Since the working areas of the human and the robot do not overlap, the spatial distribution of components on the table obliges the team to coordinate and synchronize handing-over sequences. In addition, some assembly steps require that one co-worker helps the other by fixating a part in a certain position. It is further assumed that both teammates know the construction plan and keep track of the subtasks which have been already completed by the team. The prior knowledge about the sequential execution of the assembly work is represented in the connectivity between the two layers of CSGL encoding already achieved and still to be accomplished assembly steps. Since the sequential order of tasks execution is not unique, at each stage of the construction the execution of several subtasks may be simultaneously possible.
The humanoid robot ARoS used in the experiments has been built in our lab. It consists of a stationary torus on which a 7 DOFs AMTEC arm (Schunk GmbH) with a 3-fingers dexterous gripper (Barrett Technology Inc.) and a stereo camera head are mounted. A speech synthesizer (Microsoft Speech SDK 5.1) allows the robot to communicate the result of its goal inference and decision making processes to the human user (Bicho et al., 2010).
The information about object class, position and pose is provided by the vision system. The object recognition combines color-based segmentation with template matching derived from earlier learning examples (Westphal et al., 2008). The same technique is also used for the classification of object-directed, static hand postures such as grasping and communicative gestures such as pointing.
The selection of a specific complementary behavior in AEL has to be translated into a collision-free arm and hand trajectory. As an important constraint for efficient joint action coordination, the robotics motion should be perceived by the user as smooth and goal-directed To achieve realistic temporal motor behaviors like reaching, gasping and manipulating objects we apply a global planning technique in posture space. It is formalized as a nonlinear optimization problem and allows us to integrate constraints obtained from human reaching and grasping movements such as for instance bell-shaped velocity profiles of the joints (for details see Costa e Silva et al. (2011)).

5 Results

In the following we illustrate the coordination of decisions and actions between the human and the robot in the joint assembly task by presenting video snapshots of the interactions and the associated neuronal population representations in the model network. In the examples shown, we focus for simplicity on the initial phase of the construction to explain from the perspective of the robot the impact of action observation on action selection in varying context[1]. As summarized in Table 1, there are 9 possible goal-directed sequences and communicative gestures that distinct populations in AEL and ASL represent.
Numerical values for the Joint Action Model parameters can be found in doi:10.1016/j.humov.2010.0812 (Bicho et al., 2011a).

Table 1 Goal-direct sequences and communicative gestures

Action	Sequence of motor primitives	Short description
A_1	reach wheel → grasp → attach	attach wheel
A_2	reach wheel → grasp → handover	give wheel
A_3	reach hand → grasp wheel → attach	receive wheel to attach
A_4	reach nut → grasp → attach	attach nut
A_5	reach nut → grasp → handover	give nut
A_6	reach hand → grasp nut → attach	receive nut to attach
A_7	hold out hand	request piece
A_8	point to wheel	point to wheel
A_9	point to nut	point to nut

At any point of time of the human-robot interaction only a few of these action alternatives are simultaneously possible, that is, are supported by input from connected populations. Figure 3 illustrates the competition between action alternatives in AEL and the decisions linked to overt behavior of the robot[2]. It is important to notice, however, that the competition process in ASL and AEL also works for more complex scenarios with a larger set of possible complementary behaviors (e.g., a household scenario Pinheiro et al. (2010), full construction of the 'toy vehicle' Bicho et al. (2011b)). The number of competing action representations only affects the time it takes to stabilize a suprathreshold activation pattern representing a decision (Erlhagen and Schöner, 2002).

——— Insert Figure 3 around here ———

[1] but see `http://www.youtube.com/watch?v=A0qemfXnWiE` for a video with the complete construction task

[2] video of the human-robot interactions depicted in Fig. 3 can be found in `http://dei-s1.dei.uminho.pt/pessoas/estela/Videos/JAST/Video_Fig4_Aros_Human_Toy_Vehicle.mpg`

5.1 Selection Based on an Anticipatory Model of Action Observation and Shared Task Knowledge

A cornerstone of fluent human social interactions is the ability to predict the outcomes of others' action sequences. It allows individuals to prepare actions in responses to events in the environment that will occur only a considerable time ahead. For a robot that is supposed to assist a human user in a shared task, a goal inference capacity should be used to select an action that best serves the user's future needs. But even if the human co-worker hesitates and does not show any overt behavior, a fluent team performance requires that the robot is able to take initiative and to select an action in accordance with the shared task knowledge.

These cognitive capacities are tested in the experiment depicted in Fig. 4 (video snapshots) and Fig. 5 (field activities). The experiment starts by placing the platform on the table. The vision input updates the task representation in CSGL and the activity of two populations representing the possible subgoals of attaching the wheels become suprathreshold. Initially, the two wheels are located in the working area of the human while the two nuts are located in the workspace of the robot. As shown in snapshots S1-S2 (Fig. 4), the human reaches and grasps a wheel. At the moment of the grasping, ARoS anticipates that the co-actor's motor intention is to mount the wheel on his side. It immediately decides to reach for a nut to hold it out for the human since according to the assembly plan it is the component that he will need next.

——— Insert Figure 4 around here ———

——— Insert Figure 5 around here ———

The capacity to infer the goal of the user at the time of grasping is possible because of the way in which the partner grasps an object conveys information about what he intends do with it. The robot has sequences of motor primitives in its motor repertoire that associate the type of grasping with specific final goals. A grasping from above is used to attach a wheel to the axle whereas using a side grip is the most comfortable and secure way to hand the wheel over to the co-actor. The observation of an above grip (represented in the AOL) together with information about the currently active subgoal (attach wheel on the user's side in CSGL) trigger an activation peak in ASL that represents the simulation of the corresponding 'reaching-grasping-inserting' chain (see panel a in Fig. 5, time interval T0-T1), which automatically activates the underlying goal, 'insert wheel', in the intention layer (see panel b in Fig. 5, time interval T0-T1; see also snapshot S1 in Fig. 4). Whenever the activation pattern in IL rises above threshold it initiates a dynamic updating process in the second layer of CSGL, which represents the next possible subgoal(s) for the team (see panel c in Fig. 5; see also snapshot S2 in Fig. 4, time interval T0-T1). The shared task representation allows the robot to select a complementary action that serves the user's future goal of fixing the wheel with a nut, i.e. the evolving activation pattern

in AEL (panel d in Fig. 5, time interval T0-T1) reflects the decision to ’give a nut’ to the human.
Since the robot has no wheel in its working area, an alternative decision would be to request a wheel from the user to attach it on its side of the platform. The robot’s choice to first serve the human is the result of slight differences in the input strength from populations in CSGL to associated action representations in AEL. These differences favor the execution of the user’s subtasks over the subtasks that are under the control of the robot.

However, as illustrated in snapshot S3 (Fig. 4), in this experiment the human does not attach the wheel. Instead he places the wheel back on the table, then hesitates and does not show any object-directed action. As a consequence, no suprathreshold activation exists at that time in ASL (see panel a, Fig. 5, time interval T1-T2) and activity below threshold in IL indicates that the robot has currently not attributed any action goal to the co-actor (see panel b, Fig. 5, time interval T1-T2). The robot now takes initiative and decides to request a wheel to mount it on its side of the platform (snapshot S4, Fig. 4). This change in decision is possible because the population representing the previously selected (but not yet executed) behavior to transfer a nut is not supported anymore by input from IL. On the other hand, information about currently possible subgoals and the location of parts in the two working areas create sufficiently strong input to AEL to trigger a self-stabilized activation of the population representing the ’request-wheel’ gesture (panel d, Fig. 5, time interval T1-T2).

Subsequently, the human grasps the wheel with a side grip (snapshot S5, Fig. 4). This information coded in AOL (not shown) together with information about currently active subgoals trigger a bump in ASL that represents the simulation of the corresponding ’reach-grasp-handover’ chain (panel a, Fig. 5, time interval T2-T3), which in turn automatically activates the underlying goal representation ’give wheel’ in IL (panel b in Fig. 5, time interval T2-T3). The evolving suprathreshold activation in AEL (panel b, Fig. 5, time interval T2-T3) shows the robot’s decision to receive the wheel and attach it (see also snapshots S6-S7 in Fig. 4). When the robot has attached the wheel, the vision input updates the task representations in CSGL and a new bump encoding the subsequent subgoal ’insert nut on robot’s side’ evolves (Fig. **??**, end of time interval time T2-T3). The second possible subgoal ’insert wheel on user’s side’ remains active.

Next, the user grasps again a wheel from above, ARoS predicts as before that the user will attach the wheel on his side (panel b in Fig. 5, time interval T3-T4) and decides to hand over a nut to fix the wheel (snapshots S8-S9 in Fig. 4; see panel d in Fig. 5, time interval T3-T4). Note that an alternative decision in AEL could be to ’grasp and attach a nut on the robot’s side’. The input from OML (not shown) indicating that the two nuts are located in the workspace of the robot together with the input from CSGL support the two action alternatives in AEL. As explained above, the decision process appears to be biased toward serving the human first due to the difference in input strengths from suprathreshold population activity in

CSGL. As can be seen in the snapshots S9-S11 (Fig. 4), the user attaches the wheel, and subsequently grasps the nut from the robot's hand to plug it on the axle. As the vision system detects the change in the target object, the representations of already achieved subgoals in the memory layer of CSGL are updated accordingly and the subgoal 'insert nut on robot's side' becomes active (not shown). As a consequence, a bump in AEL evolves that represents the decision of the robot to grasp and attach a nut on its side of the platform (see panel d in Fig. 5, time interval T4-T5). The overt robot behavior is depicted in snapshots S12-S14 (Fig. 4).

5.2 *Understanding Partially Occluded Actions*

In the previous example, we have seen that the robot could infer through motor simulation the co-actor's motor intention from the way the object is grasped. But what happens when the robot cannot directly observe the hand-object interaction? In natural environments with multiple objects and occluding surfaces this is a common scenario. The capacity to discern the user's motor intention and to select an appropriate complementary behavior should of course not be disrupted by missing information about the grip type used. The firing of mirror neurons in similar occluder paradigms sugge~~st~~ that working memory about objects in the scene and shared task information about what the user should do in a specific situation may sustain the motor simulation process. This is illustrated in the following interaction scenario in which only the reaching part of the user's action sequence can be observed (see Fig. 6).

——— Insert Figure 6 around here ———

In this experiment, one wheel and the two nuts are located within the working area of the robot while the second wheel is located in the user's workspace. Initially all objects are visible for the robot and their locations can thus be memorized in OML. Then a box is introduced into the scene. The robot sees the user's hand disappearing behind the occluding surface but remembers that there is a wheel behind it. Figure 7[3] illustrates the goal inference mechanism in this situation.

——— Insert Figure 7 around here ———

——— Insert Figure 8 around here ———

The corresponding population in AOL (not shown) codes only the reaching behavior. The currently possible subgoals represented in CSGL are 'insert wheel on user's side' and 'insert wheel on robot's side' (panel b in Fig. 7). The inputs from AOL and CSGL to ASL thus pre-activate the representations of two competing action

[3] for the video see `http://www.youtube.com/watch?v=7t5DLgH4DeQ`

chains associated with two possible motor intentions. The additional input necessary for goal inference comes from the information about the memorized location of the wheels in the two workspaces represented in the OML (see panel a in Fig. 7). These inputs triggers the evolution of a self-stabilized activation peak in ASL representing the action sequence 'reach wheel-grasp-insert' (see panel c in Fig. 7; see also snapshot S2 in Fig. 6). This suprathreshold activation in turn induces the evolution of a bump in IL representing the inferred goal of the human to insert the wheel (see panel a in Fig. 8). Input from IL triggers a dynamic updating process in the second layer of the CSGL, representing the next possible subgoal(s) for the user (see panel b in Fig. 8). This allows the robot, as explained in the previous example, to select a complementary action that serves the user's future needs. As can be seen when comparing the pattern of localized activation that evolves in AEL, the robot decides to serve the human by grasping a nut for handing it over (see panel c in Fig. 8 and snapshots S3-S5 in Fig. 6).

Note that the simplification for the current robotics work to represent an entire action sequence like reaching-grasping-attaching in a single population does not affect the mechanisms supporting the simulation of partially occlude actions in ASL. A chain of coupled populations of mirror neurons representing individual motor acts (Fogassi et al., 2005) may become sequentially activated above threshold by assuming that all individual population of the chain are pre-activated by input from OML and CSGL and the initial "reaching" population gets additional input from the corresponding neuronal pool in the action observation layer (Erlhagen et al., 2007).

6 Discussion

This work showed that dynamic neural fields provide a powerful theoretical framework for designing autonomous robots able to naturally interact with humans in challenging real-word environments. Flexible and intelligent robot behavior in a social context cannot be purely explained by a stimulus-reaction paradigm in which the system merely maps in a pre-determined manner current environmental inputs onto overt behavior. Dynamic neural fields explain the emergence of persistent neural activation patterns that allows a cognitive agent to initiate and organize behavior informed by past sensory experience, anticipated future environmental inputs and distal behavioral goals. The DNF architecture for joint action reflects the notion that cognitive representations, that is, all items of memory and knowledge consist of distributed, interactive, and overlapping networks of cortical populations ("cognit" from Fuster (2006)). Network neurons showing suprathreshold activity are participating in the selection of actions and their associated consequences. Since the decision-making normally involves multiple, distributed representations of potential actions that compete for expression in overt performance, the robot's goal-directed behavior is continuously updated for the current environmental context. Important for decision making in a collaborative setting, inferring others' goals from their behavior is realized by internal motor simulation based on the activation of the same joint representations of actions and their environmental effects ("mirror mechanism", Rizzolatti and Sinigaglia (2010)). Through this automatic motor resonance process, the observer becomes aligned with the co-actor in terms of actions and goals. This alignment allows the robot to adjust its behavior without explicit communication to those of the human co-actor in space and time (for an integration of verbal communication in the DNF architecture see (Bicho et al., 2010)).

The implementation of aspects of real-time social cognition in a robot based on continuously changing patterns of neuronal activity in a distributed, interactive network strongly contrasts with traditional AI approaches. They realize the underlying cognitive processes as the manipulation of discrete symbols that are qualitatively distinct and entirely separated from sensory and motor information. We do not deny that the sequence of decisions shown in our robotics experiments could be implemented by symbolic planning as well. In fact, similar joint assembly tasks have been used in the past to test AI-style control architectures for human-robot interactions (Alami et al., 2005; Hoffman and Breazeal, 2007; Steil et al., 2004). Typically, these architectures include a dedicated module that organize the high-level task of intention coordination using rule-based logic. However, the additional planning step which is needed to link the representation of every high-level decision to the level of action preparation for the robot's actuators greatly reduces the efficiency of those representations. This makes it hard or even impossible to achieve the impressive flexibility and fluency of human team performance.

In the experiments reported here, the robot-human team executed the individual assembly steps without errors and in the correct temporal order. It is important to keep in mind, however, that decisions based on noisy or incomplete sensory information and anticipated environmental inputs are fallible. It is thus no surprise that execution and prediction errors occur with some probability in complex real-world scenarios such as the joint assembly task. To work efficiently as a team, it is important that these errors are detected and compensated by one or both team members before success is compromised. Neurophysiological and behavioral findings suggest that similar neural mechanisms are involved in monitoring one's own and other's task performance (Sebanz et al., 2006) We have described in detail elsewhere how the basic DNF model of joint action coordination can be extended to include also an action monitoring function (Bicho et al., 2011b). The key idea is that specific populations integrate activity from connected neural pools or external sensory signals that carry the conflicting information. For instance, the user might want to transfer a nut to the robot but a nut has been already attached at the robot's construction side. To detect the conflict between the inferred intention of the user and the state of the construction it is sufficient to postulate that input from IL and CSGL may drive the target population beyond threshold. This suprathreshold activity may then produce (inhibitory) biasing effects for the competition between action representations in AEL. In the example, the prepotent complementary behavior of receiving the nut has to be suppressed to favor a correct response like a communicative pointing at the attached object. As integral part of the distributed network, the action monitoring thus provides just another input to the dynamic action selection process.

The applications in the domain of cognitive robotics provide new challenges for the theoretical analysis of dynamic neural fields. Most current mathematical studies are exclusively concerned with the existence and stability of characteristics patterns like bumps or traveling waves (Coombes, 2005). They do not address the spatio-temporal properties that external inputs must satisfy to generate those patterns when applied to a field at rest or in a pre-activated state. For instance, multi-bump solutions that we and others apply as a memory model for multiple items or sequential events (Ferreira et al., 2011) are known to exist when a coupling function with oscillatory decay is used (Laing et al., 2002). From an application point of view, analyzing the spatial properties of the inputs (e.g., width, relative distance etc.) that may generate multi-bump solutions when they are presented simultaneously or in sequential order is of highest importance (Ferreira , Erlhagen and Bicho, in preparation).

The present robotics implementations with hand-coded inputs from connected populations are based on the seminal analytical studies of Amari and co-workers on the formation of patterns with stationary localized stimuli. For the robotics domain, it would be highly desirable to combine the field dynamics with a learning dynamics that would allow us to establish the inter-field connections in the distributed network during training and practice. According to the principle first enunciated by Hebb (1949), memory is formed by associative synaptic modulations of connections between neuronal assemblies simultaneously excited. Important for cognitive control, persistent population activity allows the learning system to establish associations between transient events separated in time. In previous simulation studies, we have shown for instance that a rate-based Hebbian learning rule (for review of mathematical formulations see Gerstner and Kistler (2002)) can be applied to establish the goal-directed mappings for action simulation in the mirror circuit (Erlhagen et al., 2006, 2007). A more rigorous understanding of the field dynamics with the weighted, self-stabilized activity from connected populations as non-stationary input would be an important contribution for the design of an autonomous learning system.

Dynamic approaches to robotics and cognition have been often criticized to address mainly lower-level cognitive phenomena like sensory-motor coordination, path planning or perception and not the high-level cognitive capacities which are characteristics of human beings (Vernon et al., 2007). Being able to synthesize in an embodied artificial agent the cognitive demands of real-time cooperative interactions with a human co-actor shows that dynamic neural field theory provides a promising research program for bridging this gap.

Acknowledgements The present research was conducted in the context of the fp6-IST2 EU-IP Project JAST (proj. nr. 003747) and partly financed by the FCT grants POCI/V.5/A0119/2005 and CONC-REEQ/17/2001. We would like to thank Luis Louro, Emanuel Sousa, Flora Ferreira, Eliana Costa e Silva, Rui Silva and Toni Machado for their assistance during the robotic experiments.

References

Alami, R., Clodic, A., Montreuil, V., Sisbot, E. A., Chatila, R., 2005. Task planning for human-robot interaction. In: Proceedings of the 2005 Joint Conference on Smart Objects and Ambient Intelligence. ACM International Conference Proceeding Series, Vol. 121, pp. 81–85.

Amari, S., 1977. Dynamics of pattern formation in lateral-inhibitory type neural fields. Biological Cybernetics 27, 77–87.

Bicho, E., Erlhagen, W., Louro, L., Costa e Silva, E., 2011a. Neuro-cognitive mechanisms of decision making in joint action: A human-robot interaction study. Human Movement Science 30, 846–868.

Bicho, E., Erlhagen, W., Louro, L., Costa e Silva, E., Silva, R., Hipolito, N., 2011b. A dynamic field approach to goal inference, error detection and anticipatory action selection in human-robot collaboration. In: Dautenhahn, K., Saunders, J. (Eds.), New Frontiers in Human-Robot Interaction. John Benjamins, pp. 135–164.

Bicho, E., Louro, L., Erlhagen, W., 2010. Integrating verbal and nonverbal communication in a dynamic neural field architecture for human-robot interaction. Frontiers in Neurorobotics doi: 10.3389/fnbot.2010.0005.

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

Cisek, P., 2007. Cortical mechansims of action selection: the affordance competition hypothesis. Philosophical Transactions of the Royal Society B 362, 1585–1599.

Cohen, P., Levesque, H. J., 1990. Intention is choice with commitment. Artificial Intelligence 42, 213–261.

Coombes, S., 2005. Waves, bumps, and patterns in neural field theories. Biological Cybernetics 93, 91–108.

Costa e Silva, E., Costa, F., Bicho, E., Erlhagen, W., 2011. Nonlinear optimization for human-like movements of a high degree of freedom robotics arm-hand system. In: Murante, B. (Ed.), Lecture Notes in Computer Science, vol. 6794, Part III. Springer-Verlag, pp. 327–342.

Erlhagen, W., Bastian, A., Jancke, D., Riehle, A., Schöner, G., 1999. The distribution of neuronal population activation as a tool to study interaction and integration in cortical representations. Journal of Neuroscience Methods 94, 53–66.

Erlhagen, W., Bicho, E., 2006. The dynamic neural field approach to cognitive robotics. Journal of Neural Engineering 3, R36–R54.

Erlhagen, W., Mukovskiy, A., Bicho, E., 2006. A dynamic model for action understanding and goal-directed imitation. Brain Research 1083, 174–188.

Erlhagen, W., Mukovskiy, A., Chersi, F., Bicho, E., 2007. On the development of intention understanding for joint action tasks. In: 6th IEEE Int. Conf. on Development and Learning. Imperial College London, pp. 140–145.

Erlhagen, W., Schöner, G., 2002. Dynamic field theory of movement preparation. Psychological Review 109, 545–572.

Ferreira, F., Erlhagen, W., Bicho, E., 2011. A dynamic field model of ordinal and timing properties of sequential events. In: Honkela, T., Duch, W., Giorlami, M., Kaski, S. (Eds.), Lecture Notes in Computer Science 6792, Part II. Springer-Verlag, pp. 325–332.

Flanagan, J. R., Bowman, M. C., Johansson, R. S., 2006. Control strategies in object manipulation tasks. Current Opinions in Neurobiology 16, 650–659.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., Rizzolatti, G., 2005. Parietal lobe: from action organization to intention understanding. Science 308, 662–667.

Fong, T., Nourbakhsh, I., Dautenhahn, K., 2003. A survey of socially interactive robots. Robotics and Autonomous Systems 42, 143–166.

Fuster, J. M., 2006. A cognit: A network model of cortical representation. International Journal of Psychophysiology 60, 125–132.

Genovesio, A., Brasted, P. J., Wise, P., 2006. Representation of future and previous spatial goals by separate neural populations in prefrontal cortex. Journal of Neuroscience 26(27), 7305–7316.

Gerstner, W., Kistler, W. M., 2002. Mathematical formulations of Hebbian learning. Biological Cybernetics 87, 404–415.

Gibbson, J. J., 1979. The ecological approach to visual perception. Houghton Mifflin, Boston.

Gold, J. I., Shadlen, M., 2007. The neural basis of decision making. Annual Review of Neuroscience 30, 535–574.

Haazebroek, P., van Dantzig, A., Hommel, B., 2011. A computational model of perception and action for cognitive robots. Cognitive Process 12, 355–365.

Hebb, D. O., 1949. The organization of behavior. John Wiley and Sons, New York.

Hoffman, G., Breazeal, C., 2007. Cost-based anticipatory action selection for human-robot fluency. IEEE Transactions on Robotics 23, 952–961.

Kishimoto, K., Amari, S., 1979. Existence and stability of local excitations in homogeneous neural fields. J. Math. Biology 7, 303–318.

Kozma, R., 2008. Intentional systems: Review of neurodynamics, modelling, and robotics implementations. Physics of Life Reviews 5, 1–21.

Laing, C. R., Troy, W. C., Gutkin, B., Ermentrout, G. B., 2002. Multiple bumps in a neuronal model of working memory. SIAM J. Appl. Math 63, 62–97.

Levesque, H., Lakemeyer, G., 2008. Cognitive robotics. In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), Handbook of Knowledge Representation. Elsevier B. V., pp. 869–886.

Newman-Norlund, R. D., van Schie, H. T., van Zuijlen, A. M. J., Bekkering, H., 2007. The mirror neuron system is more active during complementary compared with imitative action. Nature Neuroscience 10, 817–818.

Pinheiro, M., Bicho, E., Erlhagen, W., 2010. A dynamic neural field architecture for a pro-active assistant robot. In: Proc. of 3rd IEEE/RAS-EMBS International Conference on Biomedical Robotics and Biomechatronics (IEEE BioRob 2010). pp. 777– 784.

Rizzolatti, G., Luppino, G., 2001. The cortical motor system. Neuron 31, 889–901.

Rizzolatti, G., Sinigaglia, C., 2010. The functional role of the parieto-frontal mirror circuit: interpretations and misinterpretations. Nature Reviews Neuroscience 11, 264–274.

Sandamirskaya, Y., Schöner, G., 2010. An embodied account for serial order: How instabilities drive sequence generation. Neural Networks 23, 1164–1179.

Schaal, S., 2007. The new robotics: Towards human-centered machines. HFSP Journal 1, 115–126.

Schöner, G., 2008. Dynamical systems approaches to cognition. In: Sun, R. (Ed.), The Cambridge Handbook of Computational Psychology. Cambridge University Press, pp. 101–125.

Sebanz, N., Bekkering, H., Knoblich, G., 2006. Joint action: bodies and minds moving together. Trends in Cognitive Sciences 10, 70–76.

Shepard, R. N., 1997. Toward a universal law of generalization for psychological science. Science 237, 1317–1323.

Steil, J. J., Röthling, F., Haschke, R., Ritter, H., 2004. Situated robot learning for multi-modal instruction and imitation of grasping. Robotics and Autonomous Systems 47, 129–141.

Tanji, J., Shima, K., Mushiake, H., 2007. Concept-based behavioural planning and the lateral prefrontal cortex. Trends in Cognitive Science 11, 528–534.

Trappenbeg, T., Standage, D. I., 2005. Multi-packet regions in stabilized continuous attractor networks. Neurocomputing 65(66), 617–625.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., Rizzolatti, G., 2001. I know what you are doing: A neurophysiological study. Neuron 31, 155–165.

Vernon, D., Metta, G., Sandini, G., 2007. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computtional agents. IEEE Transactions on Evolutionary Computation 2, 151–181.

Westphal, G., von der Malsburg, C., Würtz, R. P., 2008. Feature-driven emergence of model graphs for object recognition and categorization. In: Bunke, H., Kandel, A., Last, M. (Eds.), Applied Pattern Recognition, Sudies in Computational Intelligence Vol. 91. Springer Verlag, pp. 155–199.

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

Figure captions

Figure 1: Joint action model consisting of a distributed network of interconnected neural populations. It implements a flexible mapping from observed actions (layer AOL) onto complementary actions (layer AEL) taking into account the inferred action goal of the partner (layer IL), contextual cues (layer OML) and shared task knowledge (layer CSGL). The goal inference capacity is based on motor simulation (layer ASL)

Figure 2: Joint action scenario: human-robot team has to assemble a 'toy vehicle' from components that are initially distributed on a table

Figure 3: Sequence of decisions in AEL and corresponding robot behavior: (a) Temporal evolution of total input to AEL. (b) Temporal evolution of field activity showing the competition process and the sequence of decisions 'give wheel', 'insert wheel', 'point to nut' and 'insert nut'. (c) The four snapshots illustrate corresponding events of the human-robot interactions

Figure 4: Video snapshots that illustrate the capacity of the robot to infer goals, take initiative and anticipate the user's future needs

Figure 5: Field activities in layers ASL, IL, CSGL and AEL for the experiment in Fig. 4. (a) Temporal evolution of input to ASL (top) and field activity in ASL (bottom). (b) Temporal evolution of field activity in IL. (c) Updating of CSGL layer representing future subgoals based on the inferred motor intention of the user (in IL). (d) Temporal evolution of input to AEL (top) and field activity in AEL (bottom)

Figure 6: Snapshots of a video showing action understanding of partially occluded actions. Snapshot S1 shows the view of the vision system of the robot

Figure 7: Field activities for the experiment in Fig. 6. (a) Temporal evolutions of fields' activity in the OML. (b) Temporal evolution of field activity representing present possible subgoals (in CSGL). (c) Temporal evolution of input to ASL (top) and the field activity in ASL (bottom)

Figure 8: Field activities in IL, CSGL and AEL for the experiment in Fig. 6. (a) Temporal evolution of field activity in IL. (b) Updating of field representing subsequent subgoals for the user based on a prediction of his current motor intention (in CSGL). (c) The temporal evolution of input to AEL (top) and field activity in AEL (bottom)

Figures

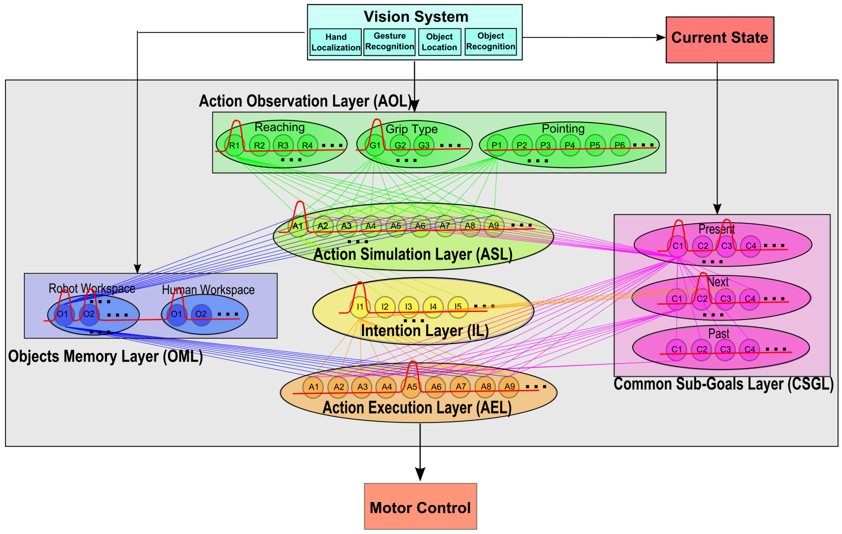


Fig. 1

Fig. 2

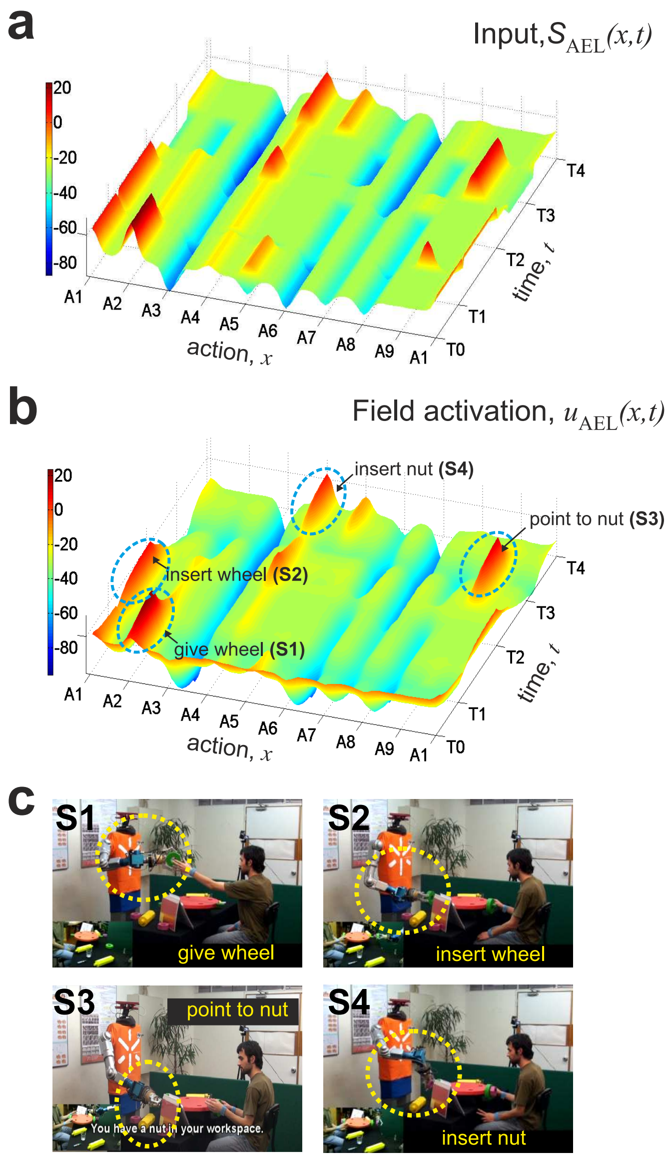
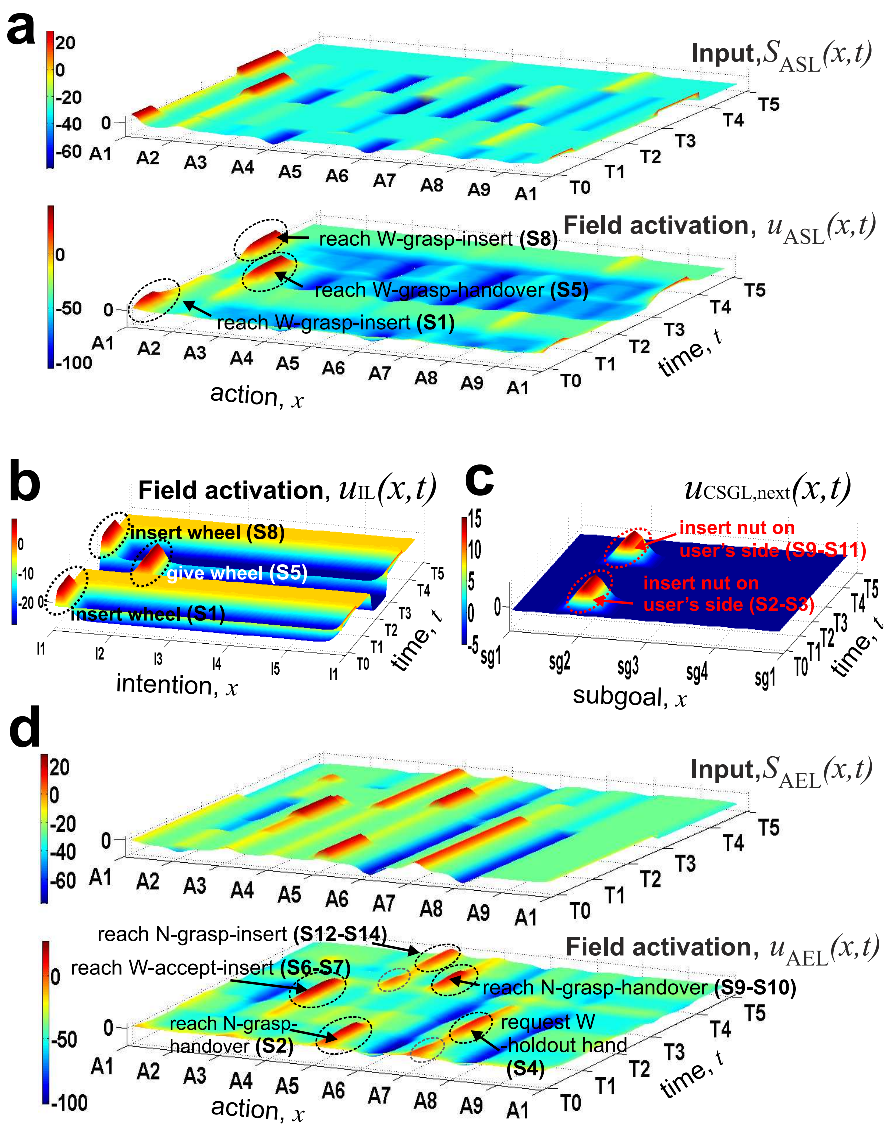

a
Input, $S_{\mathrm{AEL}}(x,t)$
20
0
-20
-40
-60
-80
A1 A2 A3 A4 A5 A6 A7 A8 A9 A1
action, x
T0 T1 T2 T3 T4
time, t
b
Field activation, $u_{\mathrm{AEL}}(x,t)$
insert nut (S4)
point to nut (S3)
insert wheel (S2)
give wheel (S1)
20
0
-20
-40
-60
-80
A1 A2 A3 A4 A5 A6 A7 A8 A9 A1
action, x
T0 T1 T2 T3 T4
time, t
c
S1
give wheel
S2
insert wheel
S3
point to nut
You have a nut in your workspace.
S4
insert nut

Fig. 3


S1
I think you are going to insert the wheel.
S2
Next you need the nut.
S3
S4
Please give me a wheel.
S5
I think you want to give me a wheel.
S6
Thank you.
S7
S8
I think you are going to insert the wheel.
S9
Next you need the nut.
S10
Here you have it.
S11
S12
I'm going to insert a nut.
S13
S14
S15
Well done! We have finished.

Fig. 4

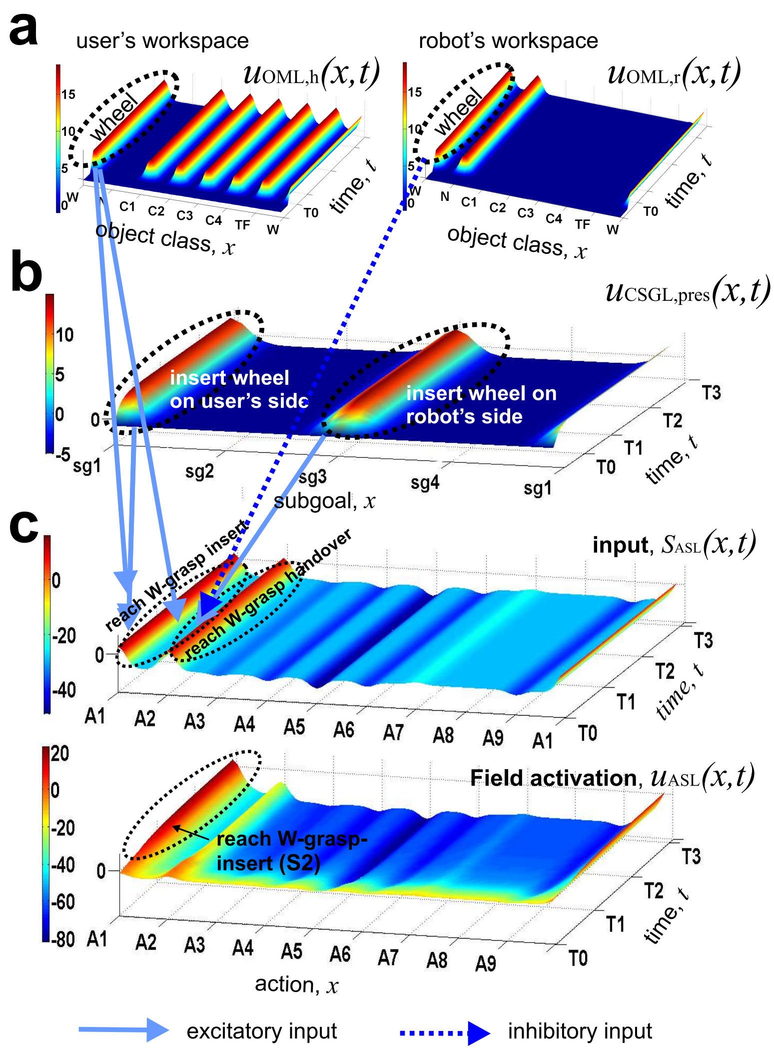


a
Input, $S_{ASL}(x,t)$
Field activation, $u_{ASL}(x,t)$
reach W-grasp-insert (S8)
reach W-grasp-handover (S5)
reach W-grasp-insert (S1)
action, x
time, t
b
Field activation, $u_{IL}(x,t)$
insert wheel (S8)
give wheel (S5)
insert wheel (S1)
intention, x
c
$u_{CSGL,next}(x,t)$
insert nut on user's side (S9-S11)
insert nut on user's side (S2-S3)
subgoal, x
d
Input, $S_{AEL}(x,t)$
Field activation, $u_{AEL}(x,t)$
reach N-grasp-insert (S12-S14)
reach W-accept-insert (S6-S7)
reach N-grasp-handover (S9-S10)
reach N-grasp-handover (S2)
request W-holdout hand (S4)

Fig. 5


S0
all objects visible
Let's start constructing the base.
S1
wheel and hand behind occluder
grip type not observable
S2
I think you are going to insert the wheel.
S3
Next you need the nut.
S4
S5
Here you have it.

Fig. 6



a
user's workspace
robot's workspace
$u_{OML,h}(x,t)$
$u_{OML,r}(x,t)$
wheel
wheel
object class, x
object class, x
time, t
time, t
b
$u_{CSGL,pres}(x,t)$
insert wheel on user's side
insert wheel on robot's side
sg1
sg2
sg3
sg4
sg1
subgoal, x
time, t
c
input, $S_{ASL}(x,t)$
reach W-grasp insert
reach W-grasp handover
time, t
Field activation, $u_{ASL}(x,t)$
reach W-grasp-insert (S2)
action, x
time, t
excitatory input
inhibitory input

Fig. 7

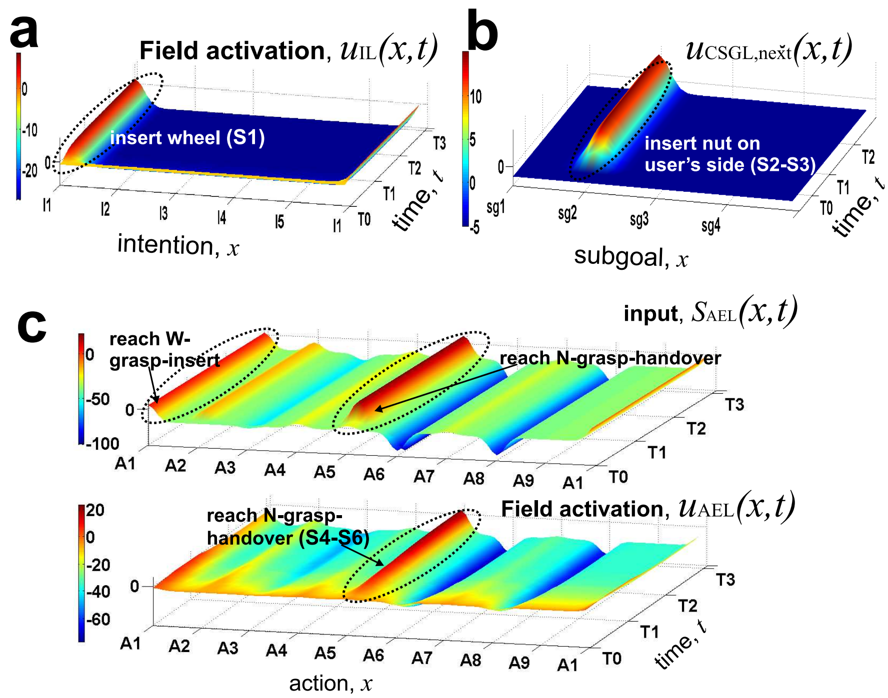

a
Field activation, $u_{IL}(x,t)$
insert wheel (S1)
I1
I2
I3
I4
I5
I1
T0
T1
T2
T3
intention, x
time, t
b
$u_{CSGL,next}(x,t)$
insert nut on
user's side (S2-S3)
sg1
sg2
sg3
sg4
subgoal, x
time, t
c
input, $S_{AEL}(x,t)$
reach W-
grasp-insert
reach N-grasp-handover
Field activation, $u_{AEL}(x,t)$
reach N-grasp-
handover (S4-S6)
A1
A2
A3
A4
A5
A6
A7
A8
A9
A1
action, x
time, t

Fig. 8