What Will You Do Next? A Cognitive Model for Understanding Others' Intentions Based on Shared Representations

Haris Dindo and Antonio Chella

University of Palermo, RoboticsLab, DICGIM, 90138 Palermo (PA), Italy
{haris.dindo,antonio.chella}@unipa.it
http://roboticslab.dicgim.unipa.it

Abstract. Goal-directed action selection is the problem of what to do next in order to progress towards goal achievement. This problem is computationally more complex in case of joint action settings where two or more agents coordinate their actions in space and time to bring about a common goal: actions performed by one agent influence the action possibilities of the other agents, and ultimately the goal achievement. While humans apparently effortlessly engage in complex joint actions, a number of questions remain to be solved to achieve similar performances in artificial agents: How agents represent and understand actions being performed by others? How this understanding influences the choice of agent's own future actions? How is the interaction process biased by prior information about the task? What is the role of more abstract cues such as others' beliefs or intentions?

In the last few years, researchers in computational neuroscience have begun investigating how control-theoretic models of individual motor control can be extended to explain various complex social phenomena, including action and intention understanding, imitation and joint action. The two cornerstones of control-theoretic models of motor control are the goal-directed nature of action and a widespread use of internal modeling. Indeed, when the control-theoretic view is applied to the realm of social interactions, it is assumed that *inverse* and *forward* internal models used in individual action planning and control are re-enacted in *simulation* in order to understand others' actions and to infer their intentions. This *motor simulation* view of social cognition has been adopted to explain a number of advanced *mindreading* abilities such as action, intention, and belief recognition, often in contrast with more classical cognitive theories - derived from rationality principles and conceptual theories of others' minds - that emphasize the dichotomy between action and perception.

Here we embrace the idea that implementing mindreading abilities is a necessary step towards a more natural collaboration between humans and robots in joint tasks. To efficiently collaborate, agents need to continuously estimate their teammates' proximal goals and distal intentions in order to choose what to do next. We present a probabilistic hierarchical architecture for joint action which takes inspiration from the idea of motor simulation above. The architecture models the casual relations between observables (e.g., observed movements) and their hidden causes

R. Shumaker (Ed.): VAMR/HCII 2013, Part I, LNCS 8021, pp. 253–266, 2013.

(e.g., action goals, intentions and beliefs) at two deeply intertwined levels: at the lowest level the same circuitry used to execute my own actions is re-enacted in simulation to infer and predict (proximal) actions performed by my interaction partner, while the highest level encodes more abstract task representations which govern each agent's observable behavior. Here we assume that the decision of what to do next can be taken by knowing 1) what the current task is and 2) what my teammate is currently doing. While these could be inferred via a costly (and inaccurate) process of inverting the generative model above, given the observed data, we will show how our organization facilitates such an inferential process by allowing agents to *share* a subset of hidden variables alleviating the need of complex inferential processes, such as explicit task allocation, or sophisticated communication strategies.

Keywords: joint action, motor simulation, shared representations, human-robot collaboration.

1 Introduction

Consider two agents (being human or artificial) collaborating on a joint task (e.g. building something together). How do they coordinate their actions without previous agreements or conventions? How do they adapt their actions during task execution? How do they achieve their goals? What are the computational mechanisms behind social interactions and joint action?

Here we argue that collaborative tasks (and social interaction problems, in general) require that interacting agents solve complex *mindreading* problems such as action and intention understanding, in parallel with motion planning and control. Indeed, recent research in social neuroscience has revealed that understanding the intentions of co-actors and predicting their next actions are fundamental for successful social interactions (cooperative or competitive) and joint actions [1,2]. In joint task such as building something together or running a dialogue, predictive mechanisms help the real-time coordination of one's own and the co-actor's actions and contribute to the success of the joint goal [3,4].

In the last years, there has been an increasing interest in joint action in the fields artificial intelligence and robotics (for a survey of related works see [5,6,7,8]) with the goal to make human-robot (or human-machine) collaboration increasingly more natural. To this aim, early researchers in AI have recognized the necessity to explicitly address the role of abstract social cues, such as intentions and beliefs, to efficiently handle teamwork problems [9]. Since then, various approaches have been built by adapting tools from symbolic reasoning [10], probabilistic decision processes [11], game theory [12] or by adopting a holistic approach based on cognitive architectures [13].

However, action understanding and prediction are hard (and often under-constrained) computational problems, and it is still unclear how humans solve them in real time while at the same time planning their complementary (or competitive) actions. It has been argued that action and intention recognition are

facilitated in joint action (but also more in general in social set-ups) because co-actors tend to automatically align (at multiple levels, of behavior and of cognitive representations), imitate each other, and share representations; in turn, this facilitates prediction, understanding, and ultimately coordination[14,15,16,17].

Here we present a computational (Bayesian) account for joint action, in which two or more agents act together so to realize a common goal. Inspired by ideas from computational neuroscience, our model describes joint action as a hierarchical phenomenon: (1) at the higher level, agents have to understand actions executed by other agents and their associated goals, and select actions that are complementary to those of the other agents or at least do not conflict with them; (2) al the lower level, agents have to coordinate their actions in real time and this requires a precise estimation of the timing and trajectories that is not necessary at the high level. Our model postulates that (shared) cognitive variables, such as beliefs and intentions, govern the activity of the motor system involved both in executing own actions and perceiving and understanding that of others via a *motor simulation* process. The following section provides a scientific background of our approach.

1.1 Background

Recognizing *what* another agent is doing and *why* (i.e., its distal intention) is extremely useful in social scenarios, both cooperative and competitive. Humans (and other animals adapted to social scenarios) are equipped with mechanisms for predicting and recognizing actions executed by others, inferring their underlying intentions, and planning actions that are complementary to them. An important constituent of the social mind of humans and monkeys is a neural mechanism for motor resonance, or the mapping observed actions into one's own motor repertoire: the *mirror system* [14]. This mechanism is part of a wide brain network that gives access to the cognitive variables (e.g., action goals and prior intentions) of another individual and permits to reconstruct the generative process that it uses to select the observed movements [16].

In this vein, it has been suggested that control-theoretic models of individual motor control can be extended to explain complex phenomena in social cognition [18,19]. The two cornerstones of control-theoretic models of motor control are the goal-directed nature of action, and the widespread use of internal modeling [20]. Indeed, when the control-theoretic view is applied to the realm of social interactions, the core scientific hypothesis is that these can be expressed through the overt and covert activity of predictive (i.e. forward) and prescriptive (i.e. inverse) internal models used in individual action planning and control [21]. In other words, an observing or interacting agent puts itself in others' shoes and and elicits its own goal-directed representations in simulation to provide an embodied explanation of others' behavior. Apparently unrelated phenomena such as motor control [22], affordance recognition [23], imitation learning [24], action understanding [25], and joint action [26] - just to name a few - can efficiently and parsimoniously be explained by the process of internal re-enactment of one's own motor apparatus: a forward model can be used as simulator of the

consequences of an action, and when paired with an inverse model a degree of discrepancy between what I observe and what I do (or just "imagine" of doing) can be produced affording better understanding of their underlying goal [21,27]. These mechanisms of *motor simulation* could act in concert with other cognitive processes such as those regulating social attention, as well as with more demanding and deliberate ones, such as those that provide a full "theory of mind" [28].

Action understanding can be related to the estimation of the (most likely) current action another agent is performing, while deeper forms of mindreading can be associated to the inference of its intentions and beliefs. According to motor theories of cognition, the same architecture used for action planning and execution can be reused for understanding actions performed by others, and their underlying intentions. In addition to these high-level problems, the low level details of action specification, prediction and adaptation are solved on-line once a motor primitive is selected. However, low-level processes can influence the choice of cognitive variables, too. Indeed, the interplay between the two levels is bidirectional: the temporal unfolding of high-level constructs biases the action recognition process which, in turn, provides necessary information to monitor the execution of the joint task itself.

From a computational viewpoint, our model of mindreading implements the idea of competition between coupled inverse and forward models [27,21], but uses approximate Bayesian inference for solving the problem. A different proposal is that of [11], in which action understanding is realized through "inverse planning" methods, and for this reason is more closely related to the idea of teleological reasoning [29] than to the idea of motor simulation that we have put forward. Our model of joint action is related to the probabilistic model of [30] in that it includes a hierarchy of representations, but it also emphasizes the formation of shared representations and their role in guiding inferential processes. Finally, our analysis is related to other initiatives that investigated the neurocognitive mechanisms that make joint action so easy [31] .

It emerges from our discussion that actions of an agent engaged in joint activities are governed by a continuous process of (joint) goal pursuing and adaptation to (1) the environment with its contextual constraints, and (2) the physical and interpersonal constraints offered by the actions of the co-actor and its abilities. The interplay of deliberate processes, which act on longer time scales, and faster processes of adaptation to the environment and the others, points to hierarchical models of action organization, with motor elements that belong to multiple levels of hierarchy (and give rise to processes that have different duration in time).

1.2 Are Shared Representations the Key for Successful Joint Actions?

Even if we assume the aforementioned hierarchical organization of action, it is currently unknown how the brain solves high- and low-level problems of joint action in real-time, given that their complexity is high even in simple scenarios [11].

We propose that co-actors do not solve interaction problems in isolation, but rather *with* the others (as well as with the environment): co-actors align their cognitive variables (beliefs, intentions and actions) and form *shared representations* (SR). We argue that what is shared during an interaction are the same representations for action (beliefs, intentions and actions) as used in individualistic action selection, performance and monitoring. For this model to work, it is not necessary that co-actors maintain separated representations for their own and another's actions, additional "we-representations", or meta-representations of what is shared. Rather, we call "shared" the subset of action representations that become aligned during interaction, being the co-actors aware of it, or not.

A first advantage of SRs is that the same cognitive variables can be used for action execution and prediction of another's actions (as well as for monitoring of the joint goal). Second, by sharing representations, an agent can help the other to understand and predict it's own actions, and to select the next action to take; although this would not be optimal from an individualistic viewpoint, it can become so if the two agents are pursuing a joint action[1].

From a computational viewpoint, shared representations help solving interaction problems in that they afford an *interactive strategy* for coordination that makes action selection and understanding easier. Put in simple terms, each agent involved in the joint action can:

1. Use motor simulation to infer what the other agent is doing (i.e., its actions) and why (up in the hierarchy of actions and intentions);
2. Infer which belief (and thus the associated sequence of intentions and actions) is the most likely one given the observed action, and 'align' it's own belief;
3. Predict what is likely to happen next by using it's own (chain of) intention and action representations, and in doing so, recognize affordances made possible (now or in the future) by the ongoing actions of the other agent;
4. Select complimentary (or successive) action by simply inferring what comes next in one's own intention and action representations (e.g., if I recognize that you are executing a certain action, I can start executing the next one in the sequence leading to the common goal);
5. While executing, lower level details are solved by other mechanisms of coordination and synchronization of action (e.g automatic entrainment, feedback, and motor simulation); in turn, as these mechanisms influence the choice of motor primitives, they have a bottom-up effect on the choice of cognitive variables;
6. When the confidence on the alignment of the joint goal is high - or when the details regarding the execution of the other agent are not essential - parts of this process can be skipped; for instance, in many circumstances co-actors can simply monitor the joint goal and use motor simulation only if an error is detected.

[1] An additional benefit of using shared representations is that, if each agent is confident that the other will facilitate it, for instance by signaling important events at the right time, then they can skip many costly mindreading and predictive processes.

We briefly mention that shared representations can be formed automatically or intentionally [32]. While in this paper we study automatic formation of shared representations, it is worth mentioning the role of *intentional* strategies that aim at influencing another's cognitive variables so as to align them to one's own. For instance, explicit communicative strategies such as the use of language, gesture, and deictics have the goal of forming or modifying shared representations. However, in [??] we focus on another - less studied - form of sensorimotor communication called *signaling*. Pushing a jointly-lifted table in a specific direction, over-articulating in noisy environment, and over-emphasizing vowels in child-directed speech are all examples of signaling. In all these examples, humans intentionally modify their action kinematics to make their goals easier to recognize. Thus, signaling acts in concert with automatic mechanisms of resonance, prediction, and imitation, especially when the context makes actions and intentions ambiguous and difficult to read. An in-depth discussion of how signaling helps joint interactions is out of the scope of the present paper (an interested reader can consult [26]).

Irrespective of how a shared representations are established, the common ground can be used as a coordination tool between two or more agents, like a blackboard in which two agents can read and write, which facilitates prediction of another's behavior by drastically reducing uncertainty, and implicitly favors the unfolding of interactive sequences of behaviors in the two agents. It emerges from our analysis that the use of shared representations changes the nature of the (high level) interaction problem from the understanding and coordination with another's actions to the active guidance of its beliefs, expectations and decisions. An agent can solve the problems of "what should I do next?" and "what will you do next?" by first inferring "what is the joint task?" and then using this information to solve the former problems. The next section provides a computational account of this process.

2 A Probabilistic Model of Joint Action

Social interaction in real world scenarios is an inherently stochastic process: perception and execution of motor acts are corrupted by noise and subject to failure, while planning of one's own acts is subordinated to the recognition of others' intentions and beliefs which are not directly observable. Furthermore, processes involved are tightly coupled (e.g. recognizing your goal-directed actions helps me updating my belief of the shared task being executed and predicting and anticipating your next steps).

We adopt the formalism of probabilistic graphical models embedding the idea of two levels of processing: at the lowest level the same circuitry used to execute my own actions are used to infer and predict the actions performed by my interaction partner via *motor simulation*, while at the highest level the two agents share action representations relative to the goals and tasks to be performed. The prior assumptions and beliefs about the joint task bias the action recognition process, while the specific motor acts confirms or disconfirms our current beliefs.

It is worth noting that two processes operate on different time scales: while the lowest level operates in real-time, providing an updated recognition of others' actions, the highest level is involved with less frequent transitions and it depends on the successful outcome of the lower levels.

In the next sections we will present our computational model focusing separately on the high- and low-level processes represented as Dynamic Bayesian Networks (DBNs). DBNs are Bayesian networks representing temporal probability models in which directed arrows depict assumptions of conditional (in)dependence between variables [33]. The general DBN model is defined by a set of N random variables $\mathbf{Y} = \{Y^{(1)}, Y^{(2)}, ..., Y^{(N)}\}$ and a pair $\{BN^p, BN^t\}$ where BN^p represents the prior $P(\mathbf{Y_1})$ and BN^t is a two-slice temporal Bayesian network which defines

$$P(\mathbf{Y_t}|\mathbf{Y_{t-1}}) = \prod_{i=1}^{N} P(Y_t^i|Pa(Y_t^i)) \tag{1}$$

where Y_t^i is the i-th node at time t and $Pa(Y_t^i)$ are the parents of Y_t^i in the graph (being in the same or previous time-slice). Usually, the variables are divided into hidden *state* variables, $\mathcal{X}$, and *observations*, $\mathcal{Z}$[2]. From the computational point of view, the task of an *inference* process is to estimate the posterior joint distribution of hidden state variables at time t, given the set of observed variables so far[3]. By marginalizing the posterior distribution it is possible to answer questions about particular variables in the network (e.g. what is the probability that a particular motor act has been executed at time t?). Next two sections provide an overview of our architecture for joint action (for a detailed description of various processes, and for an analysis of the experimental results, please consult [25,26].

2.1 Low-Level Model

The low-level model implements a *motor simulation* process that guides perceptual processing and provides action recognition capabilities. In motor simulation, it is the reenactement of one's own internal models, both inverse and forward, used for interaction that provides an understanding of what others are doing.

The entire process of action understanding can be cast into a Dynamic Bayesian Network (DBN) shown in Figure 1(a). As usual, shaded nodes represent observed variables while others are hidden and need to be estimated through the process of probabilistic inference. The model embeds the idea of motor simulation by including a probabilistic representation of forward and inverse models activation. In our representation, the process of action understanding is influenced by the following factors expressed as stochastic variables in the model (fig. 1b):

1. MP: index of the agent's own repertoire of goal-directed motor primitives; each motor primitive directly influences the activation of related forward and inverse models;

[2] By convention the observed variables are represented as shaded nodes in the network.
[3] This process is also known as filtering.

2. u: continuous control variable (e.g. forces, velocities, ...);
3. x: state (e.g. the position of the demonstrator's end-effector in an allocentric reference frame);
4. z: observation, a perceptual measurement related to the state (e.g. the perceived position of the demonstrator's end-effector on the retina).

Figure 1c shows the conditional distributions which arise in the model. The semantics of the stochastic variables, and the concrete instatiation of the conditional distributions depends on the experimental setting. Suppose we can extract the noisy measurements of the true state of the demonstrator, z_t, through some predefined perceptual process described probabilistically by the observation model $p(z_t|x_t)$. Motor primitive index variable, MP, is associated with a paired inverse-forward model, and it implicitly encodes the demonstrator's goal (in terms of the perceiver's one). The initial choice of which internal models to activate is biased by the prior probabilities (here set by the high-level network). Each paired internal model MP_t is responsible of both generating a motor control u_t, given the (hidden) state x_{t-1} (inverse model), and of predicting the next (hidden) state x_t, given the motor control u_t and the previous state x_{t-1} (forward model).

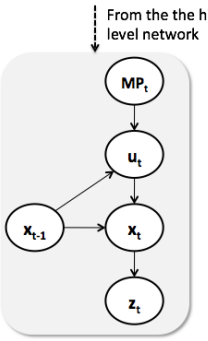
Given that in our model each goal-directed action is encoded as a coupled forward/inverse model, to predict and understand the actions performed by others it is sufficient to compute the posterior distribution over possible forward-inverse action pairs given all the observations so far, $p(MP_t|z_{1:t})$. This distribution can be obtained by marginalizing the full conditional posterior (i.e. belief) over all hidden variables in the model. Let us denote with $\mathcal{X}_t$ the set of hidden variables at time t, and with $\mathcal{Z}_t$ the set of observed variables at the same time step, the full conditional posterior can be obtained by the well-known recursive Bayesian inference schema [33]:

$$p(\mathcal{X}_t|\mathcal{Z}_{1:t}) = \eta p(\mathcal{Z}_t|\mathcal{X}_t) \cdot \int p(\mathcal{X}_t|\mathcal{X}_{t-1}) \cdot p(\mathcal{X}_{t-1}|\mathcal{Z}_{1:t-1}) d\mathcal{X}_{t-1} \tag{2}$$

where $p(\mathcal{X}_t|\mathcal{X}_{t-1})$ and $p(\mathcal{Z}_t|\mathcal{X}_t)$ are called prediction and observation models, respectively.

However, in order to compute the most likely observed action, the recursive propagation of the posterior density $p(\mathcal{X}_t|\mathcal{Z}_{1:t})$ in equation 2 is only a theoretical possibility, and in general it cannot be determined analytically. By casting the problem of action prediction and understanding in a Bayesian framework permits to adopt efficient techniques for *approximate* probabilistic inference under the constraint of limited resources. We adopt *particle filters*, a Monte Carlo technique for sequential simulation [34]. The key idea of particle filters is to represent the required posterior density function by a set of random samples with associated weights and to compute probabilistic estimates of interested quantities based on these samples and weights. Each random sample is therefore a weighted hypothesis of an internal model activation in the action prediction task, where the weight of each particle is computed according to the divergence between the predicted state of the internal model the particle belongs to and the

observed state; intuitively, severe discrepancies between predictions produced by coupled internal models and observed percepts will lead to assigning low weights to internal models less involved in explaining the current action observation. Our approach permits to solve the problem of intention recognition in real-time under the assumption that what I am observing can be adequately explained through my own internal models. The particle filter schema allows to use a multitude of internal models, for various skills and contexts, and to focus only on those able to accurately explain the current observations [25].



(a) Low-level graphical model

MP_t	goal-directed motor primitive	*discrete* $\in \{1, \ldots, N_{MP}\}$
u_t	control	*continuous*
x_t	state	*continuous*
z_t	observation	*continuous*

(b) Stochastic variables

$p(u_t \mid x_{t-1}, MP_t)$	inverse model
$p(x_t \mid x_{t-1}, u_t, MP_t)$	forward model
$p(z_t \mid x_t)$	observation model (prediction error)

(c) Probability distributions

Fig. 1. Graphical model (DBN) for action understanding based on coupled forward-inverse models; Adapted from [25]

2.2 High-Level Model

During observation of actions executed by others, motor simulation provides information that can be used to filter perceptual processing by allocating more resources (i.e., more particles in the particle filtering algorithm) to the most likely observations. This process achieves two objectives at the same time: first, it helps perceptual processing (like in Kalman filtering), and second, it permits to recognize the observed actions at the goal level by mapping them into the perceiver's repertoire of internal models.

However, in order to initialize the low-level portion of the network, we need to set the prior probability distribution over the goal-directed internal model pairs. In a joint task this distribution should be estimated by a higher-order process connected with the more abstract task representation. Some motor acts, viewed as paired forward-inverse models, are more probable at a point in time during the execution of a particular joint task. Therefore, the high-level portion of our computational model should bias the action recognition process, while at the same time providing a parsimonious way to encode the shared representations. Additionally, the interplay between the low- and high-level portions of our network shall not be unidirectional: the recognition of others' motor acts helps also

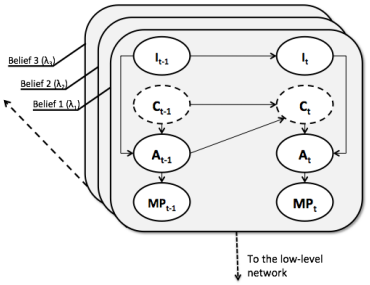
monitoring the joint act itself by revising hypotheses on the distal goal of the task in a similar vein as done in the low-level network. Here, recognized motor primitives act as observations for an abstract probabilistic representation of joint tasks and finess agents' current belief. In addition, the high-level model provides a parsimonious way to encode shared representations as explained below.

In our computational model, a joint action is influenced by three main factors: intentions, contextual information (representing the observable state of the world and its affordances) and possible actions each actor can perform given the context and intentions. The temporal evolution of these factors can be represented once again by using the formalism of probabilistic graphical models (DBN). However, each joint task requires a different motor plan, and its representation should account for possible failures in the execution. For this reason, the high level portion of our computational model includes a battery of DBNs, each one representing a possible evolution of the joint task over time (figure 2(a)). A full DBN corresponds to a belief, which intuitively encodes knowledge of "what is the task we are performing?". The stochastic variables and conditional distributions of the high-level DBN are described in figure 2(b-c).

As an example, suppose two agents (e.g. a human and a robot) have to jointly build one out of several types of towers (λ_1, λ_2, $\ldots \lambda_n$) given a set of available red and blue blocks. Each high-level network (figure 2(a)) represents a particular type of tower and can be seen as implicitly encoding the beliefs each actor has regarding the execution of the task. For instance, the tower can be made of blocks having the same color (e.g. red or blue), or of two interleaved colors (e.g. red-blue-red-blue-$\ldots$). The prior probability, $p(\lambda)$ reflects the knowledge of which tower is more probable. The variable I_t models the intention to pick and place a block of a particular color onto the tower, while the contextual variable C_t could model the availability of red and blue blocks. The action variable A_t represents the action of manipulating a particular object in the world, and it directly influences the activation of motor primitives (MP_t) used to efficiently execute the action. Motor primitives represent the observed variable and they are estimated by the low-level portion of our network at every step (see 2.1). Once an action is executed, the network models the transition to the next intention and next context through the corresponding transition probabilities.

We assume that the same set of models is shared across the two joint actors. However, their probabilistic parameters (prior, transition and observation probabilities) can be different according to individual actor's knowledge and expertise. The goal of the actors is to align their beliefs. From the probabilistic standpoint the machinery involved differs if the actor has to perform an action or if it has to recognize the action performed by another actor and update its belief. However, both computational problems have at its core the process of estimating the *likelihood* of each model given the observations.

If we denote the prior probability of a model as $P(\lambda)$, the goal is to compute the probability of the model given the set of observations so far (e.g. the likeli-



(a) High-level graphical model

I_t	intention	$discrete \in \{1, \ldots, N_I\}$
C_t	context	$discrete \in \{1, \ldots, N_C\}$
A_t	goal-directed action	$discrete \in \{1, \ldots, N_A\}$
MP_t	goal-directed motor primitive	$discrete \in \{0, \ldots, N_{MP}\}$
λ	belief	$discrete \in \{1, \ldots, N_\lambda\}$

(b) Stochastic variables

$p(I_t\|I_{t-1})$	intentional dynamics
$p(C_t\|C_{t-1}, A_{t-1})$	contextual dynamics
$p(A_t\|I_t)$	action induction
$p(U_t\|A_t)$	utility function
$p(MP_t\|A_t)$	motor primitive induction

(c) Probability distributions

Fig. 2. High-level battery of Dynamic Bayesian Networks (DBN) for joint-action. Every network in the battery is a probabilistic representation of the shared task. Adapted from [26].

hood): $P(\lambda_i|MP_{1:t})$[4]. The most plausible model is the one that maximizes the posterior probability of the model:

$$argmax_{\lambda_i} P(\lambda_i|MP_{1:t})P(\lambda_i), \forall i \in \{1, \ldots, N_\lambda\} \tag{3}$$

The likelihood is used in both action recognition and selection. In action recognition, it is used to initialize the process of motor simulation; in action selection, it is used to choose the best action to perform so that it does not lower the current likelihood. The presence of shared representations permits to describe the process in an unconventional way. Specifically, both agents use the same high-level network, in which observed and executed intentions and actions are treated on a pair, independent on who executes them. Note that the same formulation can be used to model tasks in which two agents act synchronously, such as for instance when they lift together a block, and turn-based tasks, in which one agent acts at times $t, t+2, t+4, \ldots$ and another agent acts at times $t+1, t+3, t+5, \ldots$.

The first part of the inference is the same for action observation and action selection: at each turn agents compute the likelihood of all the available models given all the observations so far (rather recognized or performed motor primitives, MP), and the belief with the highest likelihood is treated as the goal state. Action observation is then implemented as a filtering process; first, the intention I_{t+1} belonging to the current belief is predicted, which is then used to bias the recognition of MP by accordingly setting the prior probabilities needed to trigger the low-level network activation. For instance, if the system believes that the task is to build a tower made of six red blocks, it predicts that the next intention (I_{t+1}) will be to place a red block, and then it uses this information to bias the perception of actions executed by the other agent (i.e., the estimation of MP_{t+1}). In turn, the lower level affects high-level goal selection, as prediction errors drive belief revision (this is typical of hierarchical generative

[4] Likelihood computation in this network can be performed exactly by the forward-backward algorithm or approximately by the abovementioned particle filters.

models [35,36]): the recognized action is treated as an observation for the high-level network, and it is used by the observing agent to revise its current belief and eventually to align its shared representation to that of the other actor by computing the current likelihood (cf. equation 3).

Action selection is different from action observation in that MP cannot be observed (in fact, it has to be produced). Still, the process is conceptually the same: first, the intention I_{t+1} belonging to the belief with the highest likelihood is predicted; then, the most probable MP is selected for execution. For instance, if the system believes that the task is to build a tower made of six red blocks, it first predicts the most probable next intention (I_{t+1}) compatible with this belief (i.e., the intention to place a red block), then it generates an associated action (i.e., taking a specific red block), and finally an associated MP (i.e., the motor process for grasping the selected block).

3 Conclusions

Joint actions between humans and artificial agents are notoriously difficult to implement and the issue of what kind of cognitive processing is required in co-operation, coordination, and joint action is still debated. We postulate that joint actions (and social interactions in general) are heavily guided by abstract cognitive variables, such as goals, intentions and beliefs, and that the interaction itself is facilitated if interacting agents could have access to such variables. We present a computational account that allows agents to automatically align their internal representations (i.e., inferring "what task are we pursuing?" and choosing the hypothesis with higher likelihood) and then using this information in a generative scheme to both (i) decide what to do next, and (ii) predict what the other agent will do next. To cope with uncertainty, our model is developed as a two-level dynamic Bayesian network, where the lowest level implements the process of motor simulation to understand and anticipate other agent's (proximal) action intentions, while the highest level provides an abstract encoding of the task and the (distal) goals. The two levels are deeply intertwined: the temporal unfolding of high-level constructs biases the action recognition process which, in turn, provides necessary information to monitor the execution of the joint task itself. In a nutshell, our model exports the ideas from individual motor planning, control and monitoring to the realm of social interactions, by adopting the the motor view of social cognition augmented with more abstract cognitive constructs that guide the interaction.

Since any observable behavior can generally be explained by many underlying intentions and beliefs, in order to disambiguate them it is necessary to adopt costly inferential processes processes. Part of this cost can be alleviated by forming shared representations (SR) and using them as a coordination tool. Here we do not investigate the origin of shared representations; we see SR as a blackboard in which two agents can read and write and which facilitates prediction of another's behavior by drastically reducing the uncertainty of mindreading inferential processes.

Acknowledgements. The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement #231453 (HUMANOBS).

References

1. Sebanz, N., Bekkering, H., Knoblich, G.: Joint action: bodies and minds moving together. Trends Cogn. Sci. 10(2), 70–76 (2006)
2. Newman-Norlund, R.D., Noordzij, M.L., Meulenbroek, R.G., Bekkering, H.: Exploring the brain basis of joint action: Co-ordination of actions, goals and intentions. Social Neuroscience 2(1), 48–65 (2007)
3. Sebanz, N., Knoblich, G.: Prediction in joint action: What, when, and where. Topics in Cognitive Science 1, 353–367 (2009)
4. Pickering, M.J., Garrod, S.: An integrated theory of language production and comprehension. Behavioral and Brain Sciences (fort)
5. Fong, T., Thorpe, C., Baur, C.: Collaboration, dialogue, human-robot interaction. Robotics Research, 255–266 (2003)
6. Breazeal, C.: Designing sociable robots. The MIT Press (2004)
7. Sun, R.: Cognition and multi-agent interaction: From cognitive modeling to social simulation. Cambridge University Press (2005)
8. Goodrich, M.A., Schultz, A.C.: Human-robot interaction: a survey. Foundations and Trends in Human-Computer Interaction 1(3), 203–275 (2007)
9. Cohen, P.R., Levesque, H.J.: Teamwork. Nous, 487–512 (1991)
10. Breazeal, C.: Social interactions in hri: the robot view. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 34(2), 181–186 (2004)
11. Baker, C.L., Saxe, R., Tenenbaum, J.B.: Action understanding as inverse planning. Cognition 113(3), 329–349 (2009)
12. Yoshida, W., Dolan, R.J., Friston, K.J.: Game theory of mind. PLoS Comput. Biol. 4(12), e1000254+ (2008)
13. Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y.: An integrated theory of the mind. Psychological Review 111(4), 1036 (2004)
14. Rizzolatti, G., Craighero, L.: The mirror-neuron system. Annual Review of Neuroscience 27, 169–192 (2004)
15. Frith, C.D., Frith, U.: How we predict what other people are going to do. Brain Research 1079(1), 36–46 (2006)
16. Kilner, J.M., Friston, K.J., Frith, C.D.: Predictive coding: An account of the mirror neuron system. Cognitive Processing 8(3), 159–166 (2007)
17. Pezzulo, G., Candidi, M., Dindo, H., Barca, L.: Action simulation in the human brain: Twelve questions. New Ideas in Psychology (2013)
18. Grush, R.: The emulation theory of representation: motor control, imagery, and perception. Behavioral and Brain Sciences 27(3), 377–396 (2004)
19. Gardenfors, P.: Mind-reading as control theory. European Review 15(2), 223–240 (2007)
20. Wolpert, D.M., Ghahramani, Z.: Computational motor control. In: Gazzaniga, M. (ed.) The Cognitive Neurosciences III, pp. 485–494. MIT Press (2004)
21. Wolpert, D.M., Doya, K., Kawato, M.: A unifying computational framework for motor control and social interaction. Philos. Trans. R Soc. Lond. B Biol. Sci. 358(1431), 593–602 (2003)

22. Jordan, M.I., Wolpert, D.M.: Computational motor control. The Cognitive Neurosciences 601 (1999)
23. Fitzpatrick, P., Metta, G., Natale, L., Rao, S., Sandini, G.: Learning about objects through action-initial steps towards artificial cognition. In: Proceedings of IEEE International Conference on Robotics and Automation, ICRA 2003, vol. 3, pp. 3140–3145. IEEE (2003)
24. Dindo, H., Schillaci, G.: An Adaptive Probabilistic Approach to Goal-Level Imitation Learning. In: Proc. of the 2010 IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS), October 18-22, pp. 4452–4457 (2010), doi:10.1109/IROS.2010.5654440
25. Dindo, H., Zambuto, D., Pezzulo, G.: Motor simulation via coupled internal models using sequential monte carlo. In: Proceedings of IJCAI 2011, pp. 2113–2119 (2011)
26. Pezzulo, G., Dindo, H.: What should I do next? using shared representations to solve interaction problems. Experimental Brain Research 211(3), 613–630 (2011)
27. Demiris, Y., Khadhouri, B.: Hierarchical attentive multiple models for execution and recognition (hammer). Robotics and Autonomous Systems Journal 54, 361–369 (2005)
28. Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H.: Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. 28(5), 675–691 (2005); discussion 691–735
29. Csibra, G., Gergely, G.: Obsessed with goals: Functions and mechanisms of teleological interpretation of actions in humans. Acta Psychologica 124, 60–78 (2007)
30. Cuijpers, R.H., van Schie, H.T., Koppen, M., Erlhagen, W., Bekkering, H.: Goals and means in action observation: a computational approach. Neural Netw. 19(3), 311–322 (2006)
31. Vesper, C., Butterfill, S., Knoblich, G., Sebanz, N.: A minimal architecture for joint action. Neural Networks 23(8-9), 998–1003 (2010)
32. Knoblich, G., Sebanz, N.: Evolving intentions for social interaction: from entrainment to joint action. Philos. Trans. R Soc. Lond. B Biol. Sci. 363(1499), 2021–2031 (2008)
33. Murphy, K.P.: Machine Learning: A Probabilistic Perspective. MIT Press (2012)
34. Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: fifteen years later. In: Handbook of Nonlinear Filtering, pp. 656–704 (2009)
35. Friston, K.: Hierarchical models in the brain. PLoS Computational Biology 4(11), e1000211 (2008)
36. Rao, R.P., Ballard, D.H.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2(1), 79–87 (1999)