TOP-JAM: A bio-inspired topology-based model of joint attention for human-robot interaction

Hendry Ferreira Chame, Aurélie Clodic, Rachid Alami

Accepted for publication in the 2023 International Conference on Robotics and Automation (ICRA)

TOP-JAM: A bio-inspired topology-based model of joint attention for human-robot interaction†

Hendry Ferreira Chame[1,2], Aurélie Clodic[1] and Rachid Alami[1]

†This work has been partially funded by the Agence Nationale de la Recherche through the ANITI ANR-19-PI3A-0004 grant.
[1]Authors are with the team Robotics and InteractionS (RIS) at LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France, {aurelie.clodic, rachid.alami}@laas.fr.
[2]Member of the team NeuroRhythms at LORIA-CNRS, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, hendry.ferreira-chame@loria.fr.

Abstract

Coexisting with others and interacting in society implies sharing knowledge and attention about world objects, events, features, episodes, and even imagination or abstract ideas in time and space. Inspired by human phenomenological, cognitive and behavioral research, this work focuses on the study of joint attention (JA) for human-robot interaction (HRI), based on two main assumptions: a) the perception and representation of attention jointness constitute an isomorphic relation, and b) inspiration on *dynamic neural fields* (DNF) theory is a promising way to investigate contextual and non-linear spatio-temporal relations underlying attention and knowledge sharing in HRI. Taking into account the previous considerations, we propose a topology-based model for JA named TOP-JAM, which is able to represent and track in real-time JA states, from observations of behavioral data. More importantly, the model consists in a representation that can be directly understood by human beings, which conforms to robo-ethical principles in social robotics. This study evaluates computational properties of the model in simulation. Through a real experiment with the robot Pepper, the study shows that TOP-JAM is able to track JA in a triad interaction scenario.

Keywords— joint attention, neural robotics, social robotics, human-robot interaction, bio-inspired modeling.

1 Introduction

Embodied technologies are expected to be significantly present in human environments in the next few years. Thus, the field of social robotics is concerned with designing robots for natural and flexible interaction with users in public and private spaces. Among several application domains, social robots have the potential to contribute to human assistance [15], education [3], rehabilitation [19] and health care [29].

Living in society requires interacting with others, which strongly depends on the capacity of attending to oneself and to external saliency in the form of actions and events unfolding in space and time. Hence, a fundamental competence for social robots is the capacity to contextualize and represent joint attention (JA) and knowledge in human-robot interaction (HRI). This skill is crucial for the accomplishment of shared goals through *joint action* ([30], [10]), and constitutes the basis for establishing rapport in HRI (e.g. in [9]).

We argue that acceptance of social robots in society is mostly conditioned on two important factors: a) how intuitive it is for humans to interact with robots, that is, how much adaptation and learning is required from the human, and b) how trustful such technology can be.

A promising way to handle intuitiveness in HRI, is to get inspiration from human phenomenological, cognitive and behavioral studies, where JA is viewed as a cluster of cognitive processes and skills allowing individuals to focus on physical objects and others. Hence, JA is essential for coordinating, sharing, understanding and cooperating with others ([24], [36], [35], [32]).

Concerning the second factor, although some studies have suggested that people having no previous experience in HRI tend to show positive attitude and willingness to interact with robots [25], which constitutes a favorable context for their integration in society, trust in this technology is according to [11] an essential aspect to be taken into account. Thus, in order to be trusted, social robotics design and manufacturing should not only aim at achieving optimal technical performance, but also to be built according to human values, by conforming to ethical principles such as *accountability*, *responsibility* and *transparency*.

Taking into account the previous considerations, from observations on behavioral data, this work proposes a topology-based model for joint attention named TOP-JAM, which is able to represent and track in real-time JA states, based on the following assumptions:

a) The perception and representation of JA constitutes an isomorphic relation.

b) Inspiration on *dynamic neural fields* (DNF) theory is a promising way to investigate contextual and non-linear spatio-temporal relations underlying attention and knowledge sharing in HRI.

We believe that TOP-JAM is a representation general enough to be integrated to several HRI tasks and scenarios, providing valuable information for ensuring robot interaction skills such as *initiating*, *responding* or *ensuring* JA [16].

The rest of this document is organized as follows: Section 2 reviews previous research and addresses some limitations in order to justify our proposal. Section 3 presents theoretical assumptions and principles structuring our study and provides the mathematical definition of the model. Section 4 describes the methodology employed, which consisted in designing simulations and a real scenario where two subjects interact with a humanoid robot, Section 5 reports on the study's results. Finally, Section 6 presents conclusions and future perspectives.

2 Previous work

Joint attention in HRI is a vast topic of research. Some studies have investigated how humans perceive robot attention (e.g. through gaze [1], sonification [12]). Other works have considered the effects of JA on HRI, for instance in subjects diagnosed with autism spectrum disorder (ASD). In these studies, diverse response modalities are observed on the human side (e.g. brain activity in face-to-face interaction [18], behavior tracking for initiation of JA and responding to JA [5]). Commonly, the robot's behavior is pre-programmed or controlled by human operators, in order to standardize experimental conditions and to ensure that the robot executes actions in correspondence with the interaction context, which reveals how difficult it is for state-of-the-art technology in social robotics to represent and track JA states for HRI.

Several studies conducted in the Robotics and InteractionS (RIS) team have considered robot decision-making in HRI as a computational process characterized by high-level goals, where decisions should take into account the perspective of the others. Consequently, active and reactive behavior should be conditioned not only to the robot's intrinsic goals, but also to those of other agents. This perspective taking ability is considered at all decisional levels, from situation assessment based on theory of mind skills ([20, 27]), to human-aware motion and task planning ([31, 4]) and to communication management ([28]), among others. In addition, some works have dealt with more specific problems (e.g. quality of interaction [22], perspective taking for route directions [37]) where JA is an important component of the task solution. All these works lead to the definition of robotic architectures ([21]) where JA, while an important component, has not yet been studied for itself.

The previous works present some limitations which are addressed in our study. Frequently, JA in HRI is investigated in dyadic situations characterized by a task to be accomplished. The dyadic scenario in HRI is perhaps too restrictive when excluding the possibility of several humans and robots interacting together. Moreover, observations on JA tend to be coupled with abstraction on the task decision space, which prevents generalization and re-usability to different tasks scenarios. By taking into account these limitations, we propose to represent JA states for HRI inspired by research in human social cognition, which is detailed in the next section.

3 The TOP-JAM model

Theoretical Considerations. According to [32], JA involves different cognitive skills and processes constituting typologies of social attention along with corresponding levels of common knowledge. Such typology is abstracted from numerous cognitive, behavioral, and phenomenological research, and accounts for different cognitive functions or purposes in human beings.

From an experiential level of analysis, [14] describes the access possibilities to oneself and others by distinguishing three distinct perspectives. Accordingly, subjective experiences are accessible from *first-person* perspective, intersubjective or co-experiences are accessible from *second-person* perspective (see [8] and [6]), and one-way or remote observations are accessible from *third-person* perspective.

The experiential perspective adopted when relating to others is taken into account by [32] when describing social attention within a *scale of jointness* (see Table 1 and Fig. 1). This scale includes at the left extreme *individual* attention, followed by four social attention states (*monitoring*, *common*, *mutual* and *shared*). That is, a continuum is defined from no jointness at the left side to the highest degree of jointness on the right side. It is also hypothesized that shared knowledge increases when moving to the right on the scale. Our proposal for representing JA states in HRI is inspired by the scale of jointness' typology. Some typical scenarios for triadic interaction are illustrated in Fig. 2.

Table 1: Attention typology [32] for subjects q, v and object o.

Type	Description
Individual	q attends to v and o. q does not take the perspective nor connect with v's attention state.
Monitoring	q attends to v's attention to o. q knows what v is attending to.
Common	q attends to v's attention to o and to q. q and v are aware of each other's presence.
Mutual	Both q and v attend to o and get into involuntary contact (by touch or vision) while doing so.
Shared	q and v attend to o and verbally share about it.

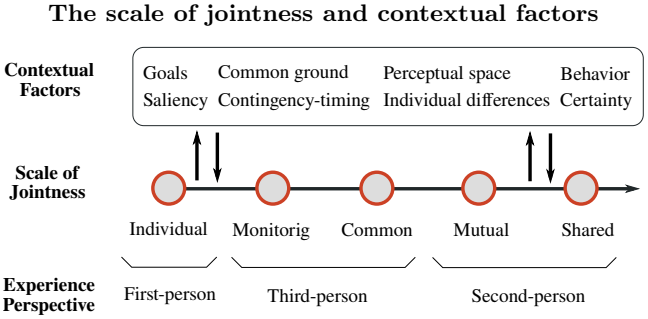

Figure 1: Contextual factors influencing joint attention and experiential perspective (inspired by [32]). Moving from left to right in the scale of jointness represents increasingly how much the other is in mind, certainty of jointness, and experienced connection felt with others.

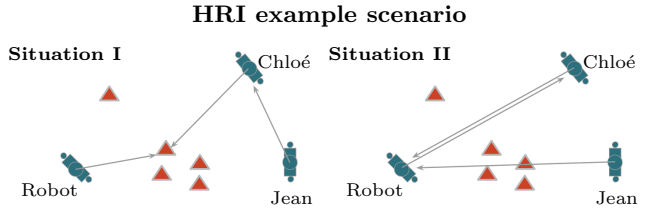

Figure 2: Two interaction situations between three actors (two humans and a robot) mediated by objects (triangles), arrows indicate the visual focus of attention. Examples of likely instantaneous JA states. For Situation I: Robot↔Chloé (*common*), Robot→Jean (*individual/monitoring*), Jean→Robot (*individual/monitoring*), Jean→Chloé (*monitoring*) and Chloé→Jean (*individual*). For Situation II: Robot↔Chloé (*mutual/shared*), Robot→Jean (*individual/monitoring*), Jean→Robot (*monitoring*), Jean↔Chloé (*common*) and Chloé→Jean (*common*).

The Mathematical Model. The mathematical formulation[1] of the model supposes HRI situations where interac-

[1] **Notation.** Matrices and vectors are represented in bold, indexes are represented as subscripts (e.g. the i^{th} element of a vector $\mathbf{a}$ is denoted $\mathbf{a}_i$). Network layers are vectors. Feature relations are denoted superscript in brackets (e.g. layer $\mathbf{y}^{[ab]}$ represents a relation between features a and b). Weight matrices are named $\mathbf{W}$. Position and orientation vectors are in 3D Cartesian space unless stated otherwise. The 3D rotation matrix about the Z axis is denoted $\mathbf{R}_z$. Unit normal directions are denoted $\hat{\mathbf{n}}$. The projection of a 3D vector $\mathbf{v}$ in the XY plane is denoted $\mathbf{v}_{\widehat{XY}}$. The dot product between vectors $\mathbf{p}$ and $\mathbf{v}$ is denoted $\mathbf{p} \cdot \mathbf{v}$.

tion is mediated by objects or environmental references characterized by intrinsic properties. Thus, it is assumed that humans and robots can recognize objects, locations, and perform and observe the following behaviors in others: a) looking-at (other's body and face, objects or references in space), b) touching (only observing would be appropriate for robots in some situations), and c) generating and understanding speech.

The experimental environment

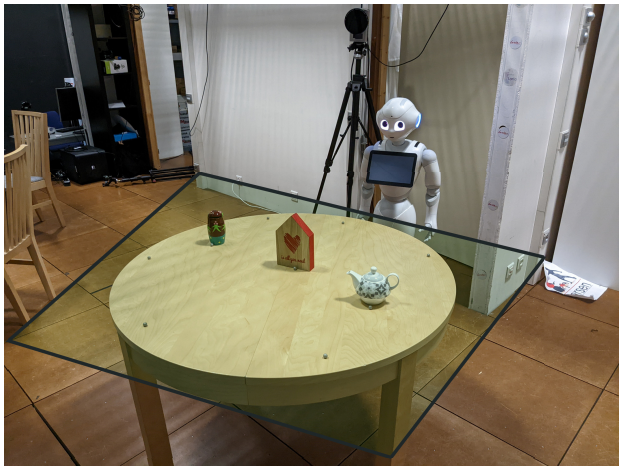

Figure 3: The shaded are illustrates the space encoded by a 2D CANN.

As illustrated in Fig. 3, an important principle underlying the model design is the definition of a topological space for JA, where spatial-temporal dynamics of attention can be represented. Hence, a neural network architecture is proposed to model JA, which is detailed next.

The network architecture. The model architecture is shown in Fig. 4. In order to compute real-time estimations of the JA states described in Table 1, the model takes as input the set of behavioral observations (input variables) described in Table 2. From these inputs, individual, social and group behavior features are computed. It is important to notice that the model does not establish a direct input-output correspondence between observations and estimations. Contrarily, it consists in a dynamical system where interaction context is taken into account and evolve according to selected temporal dynamics parametrization. In Fig. 4 two main abstractions can be observed: a) *continuous attractor neural network* (CANN) structures (for $\mathbf{I}^*$, $\mathbf{S}^*$, $\mathbf{G}^{\mathrm{a}}$, and the scale of jointness variables), and b) logical operations taking CANN states as operands (for variables $\mathbf{G}^{\mathrm{s}}$, $\mathbf{G}^{\mathrm{f}}$, $\mathbf{G}^{\mathrm{t}}$, $\mathbf{G}^{\mathrm{c}}$,$\mathbf{G}^{\mathrm{o}}$, and $\mathbf{G}^{\mathrm{b}}$).

The proposal of CANN architectures emerged within the context of *dynamic neural fields* (DNF) theory (see prominent studies such as [38], [2], and [34]), which has influenced research in robotics (e.g. [23], [17]). Inspired by the formalism proposed in [33] for modeling discrete systems, the dynamics of the continuously changing activation $\mathbf{h}_{x(t)}$ of node x at time t with preferred value v_x, is determined from the recurrent input from other neurons, external input $\mathbf{u}_{x(t)}$, and its own relaxation, such that

$$\tau \frac{\delta \mathbf{h}_{x(t)}}{\delta t} = -\mathbf{h}_{x(t)} + \rho \sum_{y} (\mathbf{W}_{xy} + \epsilon)\mathbf{f}_{y(t)} + \varsigma \mathbf{u}_{x(t)} \qquad (1)$$

where τ is the synaptic time constant, ρ and ς are scaling factors, ϵ is a global inhibition constant, and $\mathbf{f}_{y(t)}$ is the firing rate of unit y.

In Equations (5) to (9) arrow subscripts represent levels of intensity, obtained by computing the dot product between the state variable and feature extraction kernels.

TOP-JAM architectural view

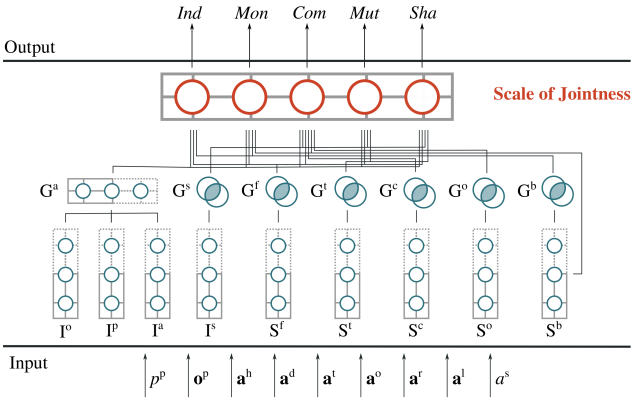

Figure 4: The TOP-JAM model. Network inputs are given in Table 2.

Taking into account the principle of local interconnections [26], the interaction strength $\mathbf{W}_{xy}$ between units x and y is defined so units representing similar states have stronger connections. Hence, Gaussian weights can be chosen, so

$$\mathbf{W}_{xy} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\mathrm{v}_y - \mathrm{v}_x)^2}{2\sigma^2}\right), \qquad (2)$$

with the parameter σ controlling the neuron range. For the case of two or higher dimension attraction spaces, multivariate Gaussian weights can be selected as well.

It is important to notice that $\mathbf{W}_{xy}$ depends on $(\mathrm{v}_y - \mathrm{v}_x)$, so recurrent interactions are *translational invariant*, which constitutes an important property of the model. Consequently, the network represents a set of stationary states continuously localized, conforming a state space manifold. The fact of disposing related JA states contiguously in the manifold representation (according to the scale of jointness, see Fig. 1) conforms to the first hypothesis of this study announced in the Introduction Section, so the perception and representation of JA states constitute an isomorphic relation where closely perceived states are represented proximal. An analogous principle has been employed in a previous work to model motivation based on self-determination theory [7].

Our model exploits an interesting possibility for JA in HRI which consists in relating topologies (representing physical space) to abstract spaces in the form of object's categorization or properties. Concretely, by establishing the weight connection $\mathbf{W}^{[\mathrm{hp}]}_{(t)}$ from Hebbian learning, it is possible to relate a CANN state $\mathbf{h}(t)$, representing focus of attention to locations in space at time t, to a one-hot encoding space representing attention to object properties $\mathbf{p}_{(t)}$, such that

$$\mathbf{p}_{(t)} = \mathbf{W}^{[\mathrm{hp}]}_{(t)} \mathbf{h}(t), \qquad (3)$$

Notice that by inverting $\mathbf{W}^{[\mathrm{hp}]}_{(t)}$ it is possible to go from object properties to locations, which can be exploited to model attention expectations or beliefs (i.e. by extending Eq. (1) with external inputs related to properties expectations or beliefs). This is inspired by *free energy principle* theory, which views organisms as proactively engaged in anticipating sensation in a generative sense from empirical priors, and minimizing free-energy as an upper-bound of surprise [13].

As previously mentioned, in the architecture shown in Fig. 4 the Group Features $\mathbf{G}^{\mathrm{s}}$, $\mathbf{G}^{\mathrm{f}}$, $\mathbf{G}^{\mathrm{t}}$, $\mathbf{G}^{\mathrm{c}}$, $\mathbf{G}^{\mathrm{o}}$, and $\mathbf{G}^{\mathrm{b}}$ are computed from logical operations, taking CANN states as operands (see the column *Input / logical operation* in Table 2). Inputs to these features consist in network states encoding g levels of a feature's intensity in uni-dimensional topological space (e.g. $g = 5$ would encode low(↓), low-mid(↘), mid(→), mid-high(↗), and high(↑) intensity levels).

Table 2: Variables definition

Input variables

ID	*Description*
p^{p}	property probability
$\mathbf{o}^{\mathrm{p}}$	object position
$\mathbf{a}^{\mathrm{h}}$	agent head position
$\mathbf{a}^{\mathrm{d}}$	agent head direction
$\mathbf{a}^{\mathrm{t}}$	agent torso position
a^{o}	agent torso yaw angle
$\mathbf{a}^{\mathrm{r}}$	agent right hand position
$\mathbf{a}^{\mathrm{l}}$	agent left hand position
a^{s}	agent speaking probability

Individual Features

ID	*Description*	*Input*
$\mathbf{I}^{\mathrm{a}}$	agent attention topology	$\mathbf{a}^{\mathrm{h}}, \mathbf{a}^{\mathrm{d}}$
$\mathbf{I}^{\mathrm{o}}$	agent attention to object	$\mathbf{I}^{\mathrm{a}}, \mathbf{o}^{\mathrm{p}}$
$\mathbf{I}^{\mathrm{p}}$	agent attention to property	$\mathbf{I}^{\mathrm{a}}, p^{\mathrm{p}}$
$\mathbf{I}^{\mathrm{s}}$	agent speaking behavior	a^{s}

Social Features (from agent q to agent v)

ID	*Description*	*Input*
$\mathbf{S}^{\mathrm{f}[qv]}$	q looks at v face	$\mathbf{a}^{\mathrm{d}[q]} \cdot \left(\mathbf{a}^{\mathrm{h}[q]} - \mathbf{a}^{\mathrm{h}[v]}\right)$
$\mathbf{S}^{\mathrm{b}[qv]}$	q looks at v body	$b_{\max}\left(\mathbf{a}^{\mathrm{d}[q]} \cdot \left(\mathbf{a}^{\mathrm{h}[q]} - \mathbf{j}^{[v]}\right)\right)$, $\mathbf{j} \in \left\{\mathbf{a}^{\mathrm{t}}, \mathbf{a}^{\mathrm{l}}, \mathbf{a}^{\mathrm{r}}\right\}$
$\mathbf{S}^{\mathrm{o}[qv]}$	q is facing v	$\left(\mathbf{R}^{z}_{(a^{\mathrm{o}[q]})}\hat{\mathbf{n}}\right)_{\widehat{\mathrm{XY}}} \cdot \left(\mathbf{a}^{\mathrm{t}[q]} - \mathbf{a}^{\mathrm{t}[v]}\right)_{\widehat{\mathrm{XY}}}$
$\mathbf{S}^{\mathrm{c}[qv]}$	q is proximal to v	$b_{\min}\left\|\mathbf{i}^{[q]} - \mathbf{j}^{[v]}\right\|$, $\mathbf{i},\mathbf{j} \in \left\{\mathbf{a}^{\mathrm{h}}, \mathbf{a}^{\mathrm{t}}, \mathbf{a}^{\mathrm{l}}, \mathbf{a}^{\mathrm{r}}\right\}$
$\mathbf{S}^{\mathrm{t}[qv]}$	q touches v	$b_{\min}\left\|\mathbf{i}^{[q]} - \mathbf{j}^{[v]}\right\|, \mathbf{i} \in \left\{\mathbf{a}^{\mathrm{l}}, \mathbf{l}^{\mathrm{r}}\right\}$, $\mathbf{j} \in \left\{\mathbf{a}^{\mathrm{h}}, \mathbf{a}^{\mathrm{t}}, \mathbf{a}^{\mathrm{l}}, \mathbf{a}^{\mathrm{r}}\right\}$

Group Features (reciprocity for agents q and v)

ID	*Description*	*Input / logical operation*
$\mathbf{G}^{\mathrm{a}[qv]}$	Topology attention	$b_{\max}\left(\mathbf{i}^{[q]} \cdot \mathbf{j}^{[v]}\right), \mathbf{i},\mathbf{j} \in \{\mathbf{I}^{\mathrm{a}}, \mathbf{I}^{\mathrm{o}}, \mathbf{I}^{\mathrm{p}}\}$
$\mathbf{G}^{\mathrm{s}[qv]}$	Talking	$\mathbf{I}^{\mathrm{s}[q]} \wedge \mathbf{I}^{\mathrm{s}[v]}$
$\mathbf{G}^{\mathrm{o}[qv]}$	Facing	$\mathbf{S}^{\mathrm{o}[qv]} \wedge \mathbf{S}^{\mathrm{o}[vq]}$
$\mathbf{G}^{\mathrm{t}[qv]}$	Physical contact	$\mathbf{S}^{\mathrm{t}[qv]} \vee \mathbf{S}^{\mathrm{t}[vq]}$
$\mathbf{G}^{\mathrm{c}[qv]}$	Proximity	$\mathbf{S}^{\mathrm{c}[qv]} \vee \mathbf{S}^{\mathrm{c}[vq]}$
$\mathbf{G}^{\mathrm{b}[qv]}$	Looking at body	$\mathbf{S}^{\mathrm{b}[qv]} \wedge \mathbf{S}^{\mathrm{b}[vq]}$
$\mathbf{G}^{\mathrm{f}[qv]}$	Looking at face	$\mathbf{S}^{\mathrm{f}[qv]} \wedge \mathbf{S}^{\mathrm{f}[vq]}$

Considerations : Individual features represent behavior related to objects and space, excluding other actors. Social Features represent behavior related to other actors. Group Features capture reciprocity in social behavior.

In order to compute logic-based features in the model, inspired by fuzzy logic research, and keeping in mind the interest of proposing a differentiable model, the operators *and* ($\wedge$) and *or* ($\vee$) are defined based on the Boltzmann operator (i.e. $b_{\min}$ and $b_{\max}$ are soft approximations for the *min* and *max* functions, respectively), such that

$$\mathbf{A} \circ \mathbf{B} = f\left(\mathbf{A}_i, \mathbf{B}_i\right), \forall i \in \{0, \dots, n-1\}, \tag{4}$$

where $f : \mathbf{A} \times \mathbf{B} \rightarrow \mathbf{C} | \mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^n$ is set $f = b_{\min}$ for $\circ = \wedge$ and $f = b_{\max}$ for $\circ = \vee$.

From Eq. (4), the external input $\mathbf{u}^{[qv]}_{x(t)}$ for a CANN representing the scale of jointness (see Fig. 1) for a pair of actors q and v in interaction is computed, such that

$$\mathbf{u}^{[qv]}_{\mathrm{ind}(t)} = \mathbf{G}^{\mathrm{a}[qv]}_{\downarrow} \wedge \mathbf{G}^{\mathrm{t}[qv]}_{\downarrow} \wedge \mathbf{S}^{\mathrm{b}[qv]}_{\downarrow} \tag{5}$$

$$\mathbf{u}^{[qv]}_{\mathrm{mon}(t)} = \mathbf{G}^{\mathrm{b}[qv]}_{\downarrow} \wedge \mathbf{G}^{\mathrm{f}[qv]}_{\rightarrow} \wedge \left(\mathbf{G}^{\mathrm{a}[qv]}_{\searrow} \vee \mathbf{S}^{\mathrm{b}[qv]}_{\uparrow}\right) \tag{6}$$

$$\mathbf{u}^{[qv]}_{\mathrm{com}(t)} = \mathbf{G}^{\mathrm{a}[qv]}_{\uparrow} \wedge \mathbf{G}^{\mathrm{s}[qv]}_{\downarrow} \wedge \mathbf{G}^{\mathrm{b}[qv]}_{\rightarrow} \wedge \mathbf{S}^{\mathrm{b}[qv]}_{\downarrow} \wedge \left(\mathbf{G}^{\mathrm{o}[qv]}_{\rightarrow} \vee \mathbf{G}^{\mathrm{c}[qv]}_{\uparrow}\right) \tag{7}$$

$$\mathbf{u}^{[qv]}_{\mathrm{mut}(t)} = \mathbf{G}^{\mathrm{a}[qv]}_{\rightarrow} \wedge \mathbf{G}^{\mathrm{s}[qv]}_{\downarrow} \wedge \left(\mathbf{G}^{\mathrm{f}[qv]}_{\uparrow} \vee \mathbf{G}^{\mathrm{t}[qv]}_{\uparrow}\right) \tag{8}$$

$$\mathbf{u}^{[qv]}_{\mathrm{sha}(t)} = \mathbf{G}^{\mathrm{a}[qv]}_{\rightarrow} \wedge \mathbf{G}^{\mathrm{s}[qv]}_{\uparrow} \wedge \left(\mathbf{G}^{\mathrm{f}[qv]}_{\uparrow} \vee \mathbf{G}^{\mathrm{t}[qv]}_{\uparrow}\right) \tag{9}$$

Finally, in order to obtain probability distributions from CANN states, normalization or the softmax function is applied to the state variable $\mathbf{h}_{(t)}$ (see Eq. (1)).

4 Methodology

Materials and Resources. The hardware components included a personal computer Dell Precision 7560, with 62 GB RAM memory, 11$^{\mathrm{th}}$ Generation Intel® Core™ i9-11950H @ 2.60GHz × 16, and graphic card NVIDIA RTX A4000 (although the program execution did not directly use GPU resources). Data was acquired through the Qualisys Motion Capture System (Track Manager version 2020.3). The project counted on a humanoid robot Pepper, manufactured by Softbank Robotics.

The software components were implemented in Python programming language version 3.8.10. These programs were integrated into the Robot Operating System (ROS) middleware version Noetic Ninjemys. The modules were executed in the operating system Ubuntu (20.04 LTS).

The Model Implementation. The implementation of the networks defined in Eq. (1) was obtained through numerical integration, based on a Forward Euler time-stepping scheme. The full parameters of the model are provided in the project's Github repository site[2], along with information on prerequisites and instructions on how to install and execute the software.

The Simulated Scenario. A simulated scene considering situations analogous to those illustrated in Fig. 2 was designed for testing the model. Thus, three actors (a robot and two humans) were simulated interacting in a scene containing eight objects, where each possessed two features (*shape* ∈ {triangular, squares, circular} and *color* ∈ {red, green, blue}). An approximate surface of 2x2 m was simulated for the topology attractors under two resolutions ($12^2 = 144$ and $24^2 = 576$ neural units, with each neuron encoding space on a radius around 20 cm and 10 cm, respectively). It is important to mention that implementations did not focus on the aspect of software optimization, so available routines from the Python *numpy* library were directly used.

Noise was added to the simulation of input variables (see Table 2), considering low $\mathbf{X}_{\mathrm{low}}$ and high $\mathbf{X}_{\mathrm{high}}$ Cartesian coordinate limits. Thus, noise was added from multivariate Normal $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and Uniform $\mathcal{U}_{[\mathbf{X}_{\mathrm{low}}, \mathbf{X}_{\mathrm{high}}]}$ distributions, according to a random variable x following a Binomial distribution $\mathcal{B}(n = 1, p)$. That is, if x is below a given threshold Gaussian noise is added, otherwise noise following a Uniform distribution is added. We believe that this choice would allow the model to be tested in situations closer to natural environments, where observation uncertainties do not always follow the Gaussian hypothesis.

[2] The project implementation (Python 3, ROS Noetic Ninjemys) is available at: `https://github.com/henferch/TOP-JAM`

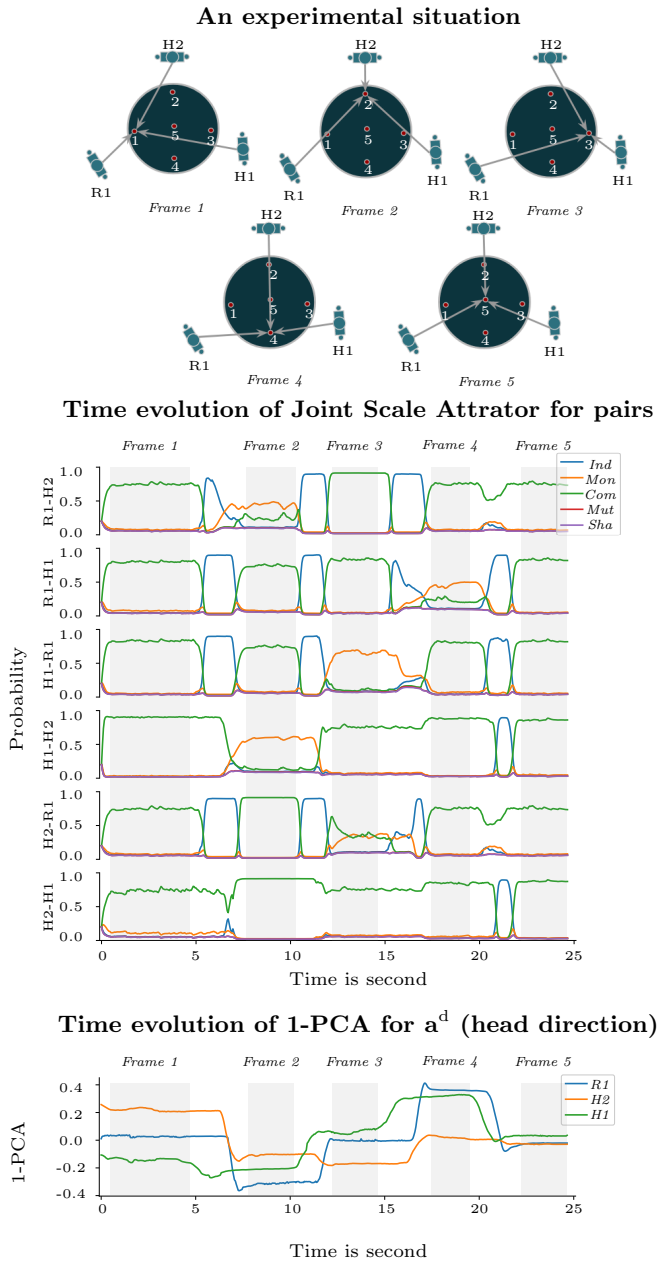

Figure 5: Above, the XY plane visualization of an experimental situation where subjects were requested to sequentially orient the head towards objects 1, 2, 3, 4 and 5 on the table (see the experiment video online). In the center, the time evolution of the softmax activation of the scale of jointness units is visualized for each pair of subjects. The time frames are shown conforming to the interval in which subjects focused simultaneously on a particular object (videos were analyzed and time coded by the experimenter). On the bottom, the time evolution of the first principal component (with explained variation of around 0.80) for input variable $\mathbf{a}^{\mathrm{d}}$ (see Table 2).

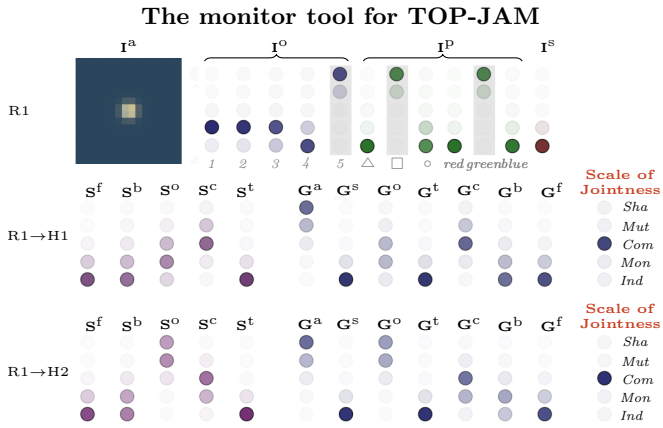

Figure 6: The monitor tool developed to follow JA states. On the left corner the matrix shows the state of the attractor representing the region on the table to which the robot R1 is directing the head (as encoded in variable $\mathbf{I}^{\mathrm{a}}$), which corresponds to a location near object 5 (see Frame 5 on the top of Fig. 5). Each circle represents the activation $\mathbf{h}_{x(t)}$ of unit x at time t (see Eq. (1)) in the respective attractor network. Attention to object 5 is for instance represented by the right-most $\mathbf{I}^{\mathrm{o}}$ network. The information encoded can be read vertically. For example, the most active unit is the top one, showing the highest intensity of attention to object 5. This object is of squared shape and green color, which is encoded by $\mathbf{I}^{\mathrm{p}}$ feature networks. The level of intensity of other features can be interpreted analogously. For the CANNs representing the scale of jointness, each unit would encode the probability of being in a node corresponding to the attention topology described in Table 1 and illustrated in Fig. 1, with external input computed conforming to Equations (5) to (9).

The Experimental Scenario. The study was conducted in the Intelligent Room facility of the lab LORIA (see Fig. 3). Two human subjects were situated more or less equidistantly from the robot around the table. Situations were analogous to those shown in Fig. 2. Subjects were asked to stare at objects and faces during some period of time. On the robot side, behavior was controlled in closed-loop. Known landmarks were attached to the back of objects, so the robot could center them in the field of view. For handling the behavior of looking at humans, the robot was trained to recognize and track the participants' faces, so it could direct the head toward participants.

Concerning the humans, reflexive markers were attached to their shoulders, torso, and the external surface of their hands. In order to track the head direction, markers were attached to eyeglass frames worn by subjects throughout the experiment. Due to differences in the body morphology of humans and the humanoid Pepper, behavior data from the robot was captured both from external markers fixed to its body to determine absolute position, combined with the computation of the direct geometric model to determine the head direction. Both the motion capture software and the robot control programs were launched in the same computer, so measurements were synchronized under the same clock.

5 Results

Concerning the aspect of computational efficiency, simulations (30 trials, 1 min duration each) showed that for a case scenario of 2 m^2 x 8 objects x 6 properties x 3 humans, selecting 144 units for the CANN models required a mean computation time of 38.48 ms with standard deviation 2.77 ms, whereas the version of 576 units required a mean computation time of 94.64 with standard deviation 2.05 ms. These figures seem reasonable for HRI in close spaces and show the feasibility of employing the model in real-time scenarios. Perhaps for the case of wide open spaces, including significantly more objects, some software optimization should be considered.

In the experiment[3], although a total of 9 cameras were employed simultaneously, data from behavior observation was available on average at 91.42% of capture time. This was mostly due to occlusions or noise affecting the Qualisys Motion Capture System. In spite of this, TOP-JAM was able to track JA states (see Fig. 5). Globally, the model was able to provide stable estimates for each pair of subjects, so units' activation on the scale of jointness were coherent with the assumptions made about the typologies defined in Table 1. Figure 6 shows how information encoded in TOP-JAM can be directly understood by human beings.

[3]The experiment video is available at: `https://youtu.be/PAIOIyw2OsQ`.

6 Conclusions

This study started from the interest in developing computational methods for improving the quality of HRI based on JA skills. The revision of previous research showed that this is actually an open and challenging topic. Some limitations were found in the literature. Firstly, representation and observation of JA is usually coupled with abstraction on the task decision space, which prevents generalization and reusability to different tasks scenarios. Secondly, most research focuses on dyadic interaction, leaving behind cases of multiple humans and robots interaction.

By taking into account these limitations, we propose a representation of JA states for HRI. Concretely, inspired by research in human social cognition, we propose a model named TOP-JAM built upon CANNs network structures, which is able to represent the dynamics and track JA for several actors interacting simultaneously. Furthermore, TOP-JAM can be conveniently integrated to existing HRI decision and control architectures as a dedicated component for the estimation of JA states. The model has the advantage of being developed as open-source technology for robotics standards such as the Robot Operating System (ROS).

In order to favor acceptance of social robotics in society, we aimed at designing a proposal in compliance with human values and ethical principles. As it has been discussed, the representation of JA through a hierarchy of CANNs and logical operations has the potential to reduce the gap between the sub-symbolic and symbolic AI worlds, in terms of being intuitive and directly readable by humans. We have shown that such representation can be computationally feasible for real-time operation in small scenarios such as the office or at home. However, more research is required for cases of wider open-space situations.

There are some aspects that could not be studied in this work, which are left for future research. Firstly, the speak recognition system was not operational during the experimental phase, so this functionality was simulated in the study. Secondly, it would be interesting to integrate observations on pointing gestures to the model, which is expected to be available in the next version of TOP-JAM. In general, different behavioral clues could be informative on attention states, depending on particular contexts and interaction modalities. In order to capture such variability, a general production rule framework is under development, having the potential to help further prototyping with TOP-JAM. Moreover, by taking advantage of the fact that resulting models are differentiable, future studies will focus on automatic selection of parameters and tuning. Finally, although behavior observations can be informative for IHR, it would be interesting to contrast them with other sources of information, such as observations on brain activity.

Acknowledgment

This research was only possible with the collaboration of colleagues from the robotics teams of both LAAS-CNRS (project ANITI) and LORIA-CNRS (project Creativ'Lab).

References

[1] Admoni, H., and Scassellati, B. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction 6*, 1 (2017), 25–63.

[2] Amari, S.-i. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics 27*, 2 (1977), 77–87.

[3] Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. Social robots for education: A review. *Science Robotics 3*, 21 (2018), eaat5954.

[4] Buisan, G., Favier, A., Mayima, A., and Alami, R. Hatp/ehda: A robot task planner anticipating and eliciting human decisions and actions. In *2022 International Conference on Robotics and Automation (ICRA)* (2022), IEEE, pp. 2818–2824.

[5] Cao, H.-L., Simut, R. E., Desmet, N., De Beir, A., Van De Perre, G., Vanderborght, B., and Vanderfaeillie, J. Robot-assisted joint attention: A comparative study between children with autism spectrum disorder and typically developing children in interaction with nao. *IEEE Access 8* (2020), 223325–223334.

[6] Chame, H. F., Ahmadi, A., and Tani, J. A hybrid human-neurorobotics approach to primary intersubjectivity via active inference. *Frontiers in psychology 11* (2020), 584869.

[7] Chame, H. F., Mota, F. P., and da Costa Botelho, S. S. A dynamic computational model of motivation based on self-determination theory and CANN. *Information Sciences 476* (2019), 319–336.

[8] Chame, H. F., and Tani, J. Cognitive and motor compliance in intentional human-robot interaction. In *2020 IEEE International Conference on Robotics and Automation (ICRA)* (2020), IEEE, pp. 11291–11297.

[9] Chevalier, P., Kompatsiari, K., Ciardo, F., and Wykowska, A. Examining joint attention with the use of humanoid robots-a new approach to study fundamental mechanisms of social cognition. *Psychonomic Bulletin & Review 27*, 2 (2020), 217–236.

[10] Clodic, A., Pacherie, E., Alami, R., and Chatila, R. Key elements for human-robot joint action. In *Sociality and normativity for robots*, R. Hakli and J. Seibt, Eds. Springer International Publishing, 2017, pp. 159–177.

[11] Dignum, V., Dignum, F., Vázquez-Salceda, J., Clodic, A., Gentile, M., Mascarenhas, S., and Augello, A. Design for values for social robot architectures. In *Robophilosophy/TRANSOR* (2018), pp. 43–52.

[12] Frid, E., and Bresin, R. Perceptual evaluation of blended sonification of mechanical robot sounds produced by emotionally expressive gestures: Augmenting consequential sounds to improve non-verbal robot communication. *International Journal of Social Robotics 14*, 2 (2022), 357–372.

[13] Friston, K. The free-energy principle: a unified brain theory? *Nature reviews neuroscience 11*, 2 (2010), 127–138.

[14] Fuchs, T. The phenomenology and development of social perspectives. *Phenomenology and the cognitive sciences 12*, 4 (2013), 655–683.

[15] Góngora Alonso, S., Hamrioui, S., de la Torre Díez, I., Motta Cruz, E., López-Coronado, M., and Franco, M. Social robots for people with aging and dementia: a systematic review of literature. *Telemedicine and e-Health 25*, 7 (2019), 533–540.

[16] Huang, C.-M., and Thomaz, A. L. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *2011 Ro-Man* (2011), IEEE, pp. 65–71.

[17] Jauffret, A., Cuperlier, N., and Gaussier, P. From grid cells and visual place cells to multimodal place cell: a new robotic architecture. *Frontiers in Neurorobotics 9* (2015), 1.

[18] Kompatsiari, K., Bossi, F., and Wykowska, A. Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. *Social cognitive and affective neuroscience 16*, 4 (2021), 383–392.

[19] Langer, A., Feingold-Polak, R., Mueller, O., Kellmeyer, P., and Levy-Tzedek, S. Trust in socially assistive robots: Considerations for use in rehabilitation. *Neuroscience & Biobehavioral Reviews 104* (2019), 231–239.

[20] Lemaignan, S., Sallami, Y., Wallbridge, C., Clodic, A., Belpaeme, T., and Alami, R. UNDERWORLDS: Cascading Situation Assessment for Robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2018)* (Madrid, Spain, Oct. 2018).

[21] Lemaignan, S., Warnier, M., Sisbot, E. A., Clodic, A., and Alami, R. Artificial Cognition for Social Human-Robot Interaction: An Implementation. *Artificial Intelligence 247* (June 2017), 45–69.

[22] Mayima, A., Clodic, A., and Alami, R. Towards robots able to measure in real-time the quality of interaction in hri contexts. *International Journal of Social Robotics 14*, 3 (2022), 713–731.

[23] Milford, M., and Wyeth, G. Persistent navigation and mapping using a biologically inspired slam system. *Int. J. Rob. Res. 29*, 9 (Aug. 2010), 1131–1153.

[24] Mundy, P., Sullivan, L., and Mastergeorge, A. M. A parallel and distributed-processing model of joint attention, social cognition and autism. *Autism research 2*, 1 (2009), 2–21.

[25] Naneva, S., Sarda Gou, M., Webb, T. L., and Prescott, T. J. A systematic review of attitudes, anxiety, acceptance, and trust towards social robots. *International Journal of Social Robotics 12*, 6 (2020), 1179–1201.

[26] Samsonovich, A., and McNaughton, B. L. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience 17*, 15 (1997), 5900–5920.

[27] Sarthou, G. Overworld: Assessing the geometry of the world for human-robot interaction. *IEEE Robotics and Automation Letters 8*, 3 (2023), 1874–1880.

[28] Sarthou, G., Buisan, G., Clodic, A., and Alami, R. Extending Referring Expression Generation through shared knowledge about past Human-Robot collaborative activity. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Prague (online), Czech Republic, Sept. 2021).

[29] Scoglio, A. A., Reilly, E. D., Gorman, J. A., and Drebing, C. E. Use of social robots in mental health and well-being research: systematic review. *Journal of medical Internet research 21*, 7 (2019), e13322.

[30] Sebanz, N., Bekkering, H., and Knoblich, G. Joint action: bodies and minds moving together. *Trends in cognitive sciences 10*, 2 (2006), 70–76.

[31] Singamaneni, P.-T., Favier, A., and Alami, R. Human-Aware Navigation Planner for Diverse Human-Robot Contexts. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Prague (online), Czech Republic, Sept. 2021).

[32] Siposova, B., and Carpenter, M. A new look at joint attention and common knowledge. *Cognition 189* (2019), 260–274.

[33] Stringer, S. M., Trappenberg, T. P., Rolls, E. T., and de Araujo, I. E. Self-organizing continuous attractor networks and path integration: one-dimensional models of head direction cells. *Network 13*, 2 (May 2002), 217–242.

[34] Taylor, J. G. Neural 'bubble' dynamics in two dimensions: foundations. *Biological cybernetics 80*, 6 (1999), 393–409.

[35] Tomasello, M., et al. Joint attention as social cognition. *Joint attention: Its origins and role in development 103130* (1995), 103–130.

[36] Trevarthen, C., Hubley, P., and Lock, A. Secondary intersubjectivity: Confidence, confiding and acts of meaning in the first year. action, gesture and symbol. the emergence of language. *A. Lock. New York: Academic* (1978), 183–229.

[37] Waldhart, J., Clodic, A., and Alami, R. Reasoning on shared visual perspective to improve route directions. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (2019), IEEE, pp. 1–8.

[38] Wilson, H. R., and Cowan, J. D. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal 12*, 1 (1972), 1–24.