2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

AEGO: Modeling Attention for HRI in Ego-Sphere Neural Networks

Hendry Ferreira Chame[1] and Rachid Alami[2]

***Abstract*— Despite important progress in recent years, social robots are still far away from showing advanced behavior for interaction and adaptation in human environments. Thus, we are interested in studying social cognition in human-robot interaction (HRI), notably in improving communication skills relying on joint attention (JA) and knowledge sharing. Since JA involves low-level cognitive processes in humans, we take into account the implications of Moravec's Paradox and focus on the aspect of knowledge representation. Inspired by 4E cognition principles, we study egocentric localization through the concept of sensory *ego-sphere*. We propose a neural network architecture named AEGO to model attention for each agent in interaction and show how to fuse information in a common representation space. From the perspective of *dynamic fields theory*, AEGO takes into account the dynamics of bottom-up and top-down modulation processes and the effects of neural excitatory and inhibitory synaptic interaction. In this work we evaluate the model in simulation and experiments with the robot Pepper in JA tasks based on proprioception, vision, rudimentary natural language and Hebbian plasticity. Results show that AEGO is convenient for HRI, allowing the human and the robot to share attention and knowledge about objects in scenarios close to everyday situations. AEGO constitutes a novel brain-inspired architecture to model attention that is suitable for multi-agent applications relying on social cognition skills, having the potential to generalize to several robotics platforms and HRI scenarios.**

I. INTRODUCTION

According to Moravec's paradox, although machines can perform tasks at adults' level of intelligence, such as inductive and deductive reasoning, they have tremendous difficulty with sensory-motor or social skills, as demonstrated by one-year-old children. Behind this paradox remains the question in artificial intelligence (AI) research of what sort of knowledge representation would be suitable for a machine to accomplish cognitive tasks, which has important philosophical implications. Thus, recent studies have contrasted the Cartesian (traditional) view of social cognition, as a process confined to the brain, to the notion of an *embodied*, *embedded*, *enacted* and *extended* process, unfolding between the brain, the body and the environment in interaction: a perspective known as *4E cognition* [17].

Inspired by 4E cognition, we believe that for social robots to leave the lab and adapt to human environments, it is crucial to provide them with forms of behavior regulation which take into account the dynamics of human low-level social processes, such as the capacity of engaging in *joint attention* (JA), and the possibility of those processes be modulated in direct interaction. Furthermore, as a multi-dimensional construct, JA involves cognitive skills which constitute forms of social attention at distinct levels of interaction and knowledge sharing [25], which must be considered in HRI.

[1]hendry.ferreira-chame@loria.fr.
LORIA-CNRS, Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France.
[2]rachid.alami@laas.fr
LAAS-CNRS, 7 Av. du Colonel Roche, 31400 Toulouse, France.

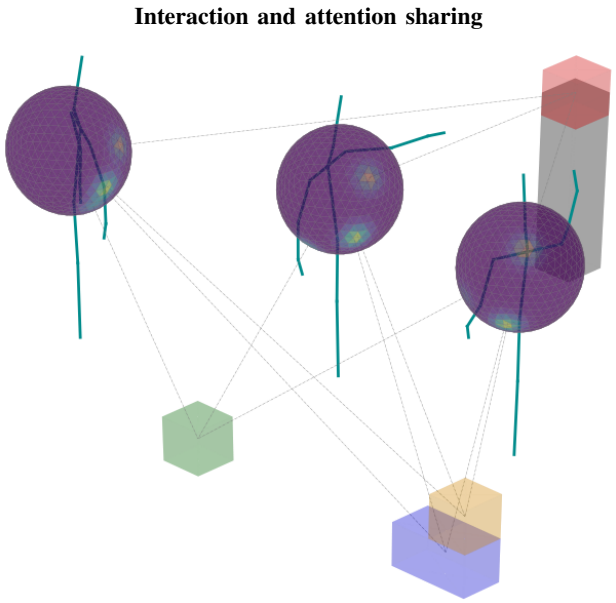

Fig. 1. Two agents are sharing attention about the red object while a third one is leaving the scene. Three other objects are present. Saliency from stimuli projection in agents' peripersonal space is represented at a pre-selection level of attention in ego-spherical localization.

Following on from a previous work that proposed tracking JA in HRI as a topology-based representation organized in a *scale of jointness* [7], here we investigate the more fundamental aspect of attention selection dynamics and how such process is an important component for the emergence of JA in HRI, under the influence of lateral connectivity, bottom-up saliency and top-down modulation activity, benefiting from online Hebbian plasticity. As shown in Fig. 1, we inquire everyday situations where agents can become interested in objects and eventually share attention and knowledge about them (e.g. asking someone for direction, commenting about a sudden noise in the environment).

Tacking into account the considerations above, we explore the concept of *ego-sphere* [1] and propose the architecture named AEGO for tracking individually agents' attention focus represented in egocentric perspective, resulting from on-board sensory acquisitions. For this, inspired by *dynamic neural fields* (DNF) theory [2], we model the attention selection process in interaction as a dynamical system. By

addressing limitations on previous research, we show how neural excitatory and inhibitory interactions allow us to study the emergence of attention selection. Moreover, we show how the model can be used to track agents' interaction with peripersonal space, which is interesting for HRI applications.

This document is organized as follows: Section II discusses previous works and how our contribution would help to advance the state of the art in the field. Section III presents the mathematical definition of the model and discusses theoretical assumptions behind it. Section IV presents the methodology which consisted in: a) studying in simulation attention selection from bottom-up and top-down modulation processes, and showing potential applications, and b) conducting an experiment with the robot Pepper for a JA task based on proprioception, vision, basic natural language and Hebbian plasticity. Section V reports on the study's results, and Section VI presents conclusions and future perspectives.

II. PREVIOUS WORK

Feature integration theory (FIT) [26] views attention as a multi-level information fusion process. According to FIT, at a pre-attention level the perceptual system receives from separate maps feature salience information (e.g., color, edges, shapes), which are lately combined at an attention selection stage. FIT has influenced several bio-inspired research (see [15], [14]), with applications in robotics (e.g. vision and autonomous navigation [24]).

According to [1] an *ego-sphere* consists of a two dimensional spherical map of the world as perceived by an observer placed at its center. This interesting idea has inspired several works in the field of robotics. A study by [19] has shown how attention and short-term memory can be modulated through saliency maps and allow the robot to explore the environment based on novelty. A work by [4] focused on intuitive HRI, including the possibility of top-down modulation of attention. The aspect of information representation has also been studied in [18], so the ego-sphere has been implemented as a storage data-base indexed by spherical tessellation mapping. Other contributions could be mentioned (e.g. [12], [16]).

To our knowledge, previous research has not explored sufficiently the aspect of interaction dynamics between locations represented in the ego-sphere, and considered at most basic forms of interaction spread between nodes. Moreover, excluding saliency map approaches (e.g. [19]), the dynamics of attention is modeled as a process governed by knowledge represented in the form of production rules, where the possibility of compositionality from low-level sensory to higher-level decision space has been of less importance.

Another limitation of previous works is considering the robot as the only agent provided with a sensory ego-sphere, so acquisitions on the human are expressed in the robot's point of view. In our opinion, this would be a too egocentric approach for HRI. We believe that tracking embodied relations between agents and objects as a dynamical system can help robot's agency and perspective-taking from a shared representation space without relying excessively on environment modeling, so improving fluidity and the emergence of JA in instantaneous interaction.

Although we are mainly interested in exploring the concept of sensory ego-sphere for JA, some contributions dealing with auditory-visual systems for HRI in allocentric perspective are related to our proposal and should be mentioned. Thus, in [9] pointing gestures are studied for domestic environments. In [23] JA is modeled in a 2D space encoding the interaction scene by combining connectionist and dynamic neural fields models. A cognitive architecture for decision-making based on JA is proposed in [10].

Our previous research also constituted relevant steps in the direction of developing the current study, which is worth mentioning. In [13] JA in HRI was studied for providing guidance in a shopping mall. Other works explored the concept of *joint action* [3] and *situation assessment* from perspective taking [20]. In [8] dynamic fields neural modeling was proposed to represent and track motivation dynamics of humans interacting with cyber-physical systems. In [6] an ego-cylindrical selection mechanism for attention was proposed for autonomous positioning with respect to objects in the environment. In [7] the model TOP-JAM was proposed for tracking JA in HRI from allocentric references.

To summarize, this work proposes to model attention inspired by FIT's hierarchical view of a process taking place at two levels (at a pre-selection and selection stages). Differently from previous works, we study egocentric attention for HRI in neural dynamic fields networks for tracking simultaneously from each participant' perspective the influence of three sources on attention selection: bottom-up stimulation, top-down modulation, and local interaction from inhibitory and excitatory synapse. The next section presents the mathematical foundations of the architecture AEGO, whereas in Section IV we show how it is suited for investigating joint attention in HRI.

III. THE MATHEMATICAL MODEL

The architectural view of AEGO is shown Fig. 2. Attention is modeled in an ego-spherical representation encoded by dynamic neural fields at a pre-attention stage, receiving stimulation from bottom-up and top-down processes, and synaptic interaction. In a subsequent stage, attention selection results from competition in synaptic interaction. The mathematical definition[1] of these layers is detailed next.

A. Attention pre-selection layer

Let the activation of the i^{th} neuron encode the dynamics of stimulation affecting a location $\mathbf{x}_i$ in 3D Cartesian coordinates at a polyhedron surface representing the agent's ego-space, such that

[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. Matrices are represented in capital letter, the colon character represents the i^{th} row or column of a matrix (e.g. $\mathbf{A}_{i:}$ for columns and $\mathbf{A}_{:i}$ for rows). Position and orientation vectors are in 3D Cartesian space. The projection of a point $\mathbf{p}$ in the ego-sphere surface is denoted $\hat{\mathbf{p}}$. A reference at the ego-sphere surface defining the preferred location for the i^{th} unit in the neural field is denoted $\mathbf{x}_i$. The y-coordinate of a point is denoted $\mathbf{p}_{\text{y}}$. The exponent notation in brackets indicates the agent to which a layer belongs (e.g. layer $\mathbf{a}^{[\text{r}]}$ belongs to the [r]obot).

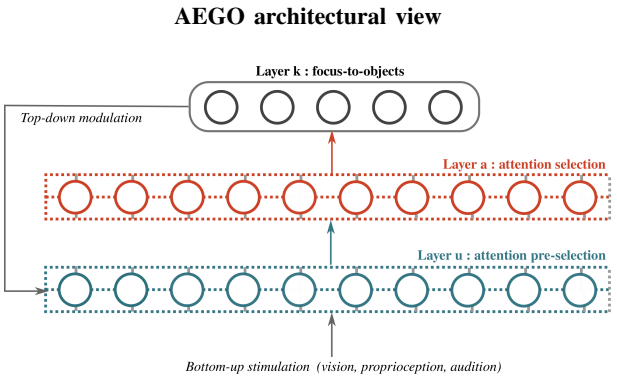

Fig. 2. In AEGO bottom-up dynamics are induced by multiple sensory systems (e.g. vision, proprioception, and audition) which excite the attention pre-selection layer $\mathbf{u}$ (see Eq. (1)) under synaptic local influences (shown in dashed lines). Attention selection is accomplished in layer $\mathbf{a}$ (see Eq. (3)) from synaptic competition. The output layer $\mathbf{k}$ (see Eq. (6)) encodes the probability of focusing on a particular object. Under top-down modulation, focused attention can influence pre-selection as a feedback process.

$$\tau_{\mathrm{u}}\dot{\mathbf{u}}_{i(t)} = -\mathbf{u}_{i(t-1)} + q_{\mathrm{u}} + \sum_j (\mathbf{U}_{ij} + \epsilon)\mathbf{u}_{i(t-1)} + \mathbf{s}_{i(t)}. \quad (1)$$

According to the principle of locality [21], the interaction strength $\mathbf{U}_{ij}$ between neurons i and j is selected so proximal locations have stronger interaction. Thus, multivariate Gaussian weights with scale factor γ_g are set such that

$$\mathbf{U}_{ij}(|\mathbf{x}_i - \mathbf{x}_j|) = \gamma_g \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right). \quad (2)$$

The term $\mathbf{s}_{i(t)}$ represents the inputs received at time instant t from bottom-up and top-down sources. Inputs for particular scenarios are discussed in Section IV. In Eq. (1) q_{u} corresponds to the activation resting state, τ_{u} is a time constant, and ϵ is a global inhibition for lateral interactions.

B. Attention selection layer

Let the activation of the i^{th} neuron represent the dynamics of attention selection at a location $\mathbf{x}_i$ in the ego-space so

$$\tau_{\mathrm{a}}\dot{\mathbf{a}}_{i(t)} = -\mathbf{a}_{i(t-1)} + q_{\mathrm{a}} + \sum_j \mathbf{A}_{ij} f\left(\mathbf{a}_{i(t-1)}, \mathbf{u}_{i(t)}\right) + \eta_{(t)}. \quad (3)$$

Inhibitory synapse has been associated with selection mechanisms [22]. Thus, we propose to model lateral interaction $\mathbf{A}_{ij}$ between neuron i and j such that

$$\mathbf{A}_{ij}(|\mathbf{x}_i - \mathbf{x}_j|) = \varphi\mathbf{U}_{ij} - 1, \quad (4)$$

with $\varphi = \max(\mathbf{U}_{i:})^{-1}$ a scaling factor. In Eq. (3) noise $\eta_{(t)} \sim \mathcal{N}(\mu_{\mathrm{a}}, \sigma^2_{\mathrm{a}})$. The activation function f is defined so

$$f\left(\mathbf{a}_{i(t-1)}, \mathbf{u}_{i(t)}\right) = \mathrm{sigmoid}\left(\gamma_{\mathrm{a}}\left(\mathbf{a}_{i(t-1)} + \gamma_{\mathrm{u}}\mathbf{u}_{i(t)}\right)\right) \quad (5)$$

being γ_{a} and γ_{u} gain constants.

C. Focus-to-objects output layer

Let the probability $\mathbf{k}_{n(t)}$ of attending to the n^{th} object be modeled as the output layer, such that

$$\mathbf{k}_{n(t)} = \mathrm{softmax}\left(\gamma_{\mathrm{k}} \sum_j \mathbf{K}_{nj(t)}\mathbf{a}_{j(t)}\right), \quad (6)$$

where $\mathbf{K}_{nj}(|\hat{\mathbf{p}}_n - \mathbf{x}_j|)$ is obtained from Eq. (2) with $\hat{\mathbf{p}}_n$ the projection of the object's center in the ego-sphere, and gain constant γ_{k}.

IV. METHODOLOGY

Three studies in simulation were designed for analyzing potential application scenarios with AEGO. An experiment was also conducted with the robot Pepper for joint attention tasks based on proprioception, vision, rudimentary natural language and Hebbian plasticity.

A. Simulations

Table I presents common parameters for Eqs. (1), (3) and (6). The network's state is obtained through numerical integration (Euler method) with time-step dt. Six objects were simulated as bottom-up saliency (see Fig. 3).

TABLE I
COMMON PARAMETERS FOR SIMULATIONS

Parameter	Value
Ego-sphere tessellation	642 vertex, 1280 faces
Ego-sphere radius	0.25 m robot, 0.3 m human
ϵ	1.0e-4
$\tau_{\mathrm{u}}, \tau_{\mathrm{a}}$	200 ms
$\boldsymbol{\Sigma}$	1.0e-3$\mathbf{I}_3$ robot, 2.0e-3$\mathbf{I}_3$ human
γ_g	6.2e-2
q_{u}	-1.0e-3
γ_{u}	2.5
q_{a}	-1.0e-4
γ_{a}	100
μ_{a}	0
σ^2_{a}	1.0e-3
γ_{k}	250
dt	50 ms

1) Focusing on named objects: we studied the possibility of attending to a specific object as modulated by top-down processes, based on the assumption that the agent is able to track and recognize objects while associating unique words for addressing them. For example, once the human says "three" the robot should recognize this word with some likelihood. In case the latter is high enough, the robot should attend to object number 3 around location $\hat{\mathbf{p}}_3$ (see Fig. 3). Thus, $\mathbf{s}_{i(t)}$ in Eq. (1) can be set so

$$\mathbf{s}_{i(t)} = \sum_n \gamma_{n(t)}\mathbf{K}_{ni(t)}. \quad (7)$$

Interest to objects is modeled through the gain $\gamma_{n(t)}$. Bottom-up saliency is set so $\gamma_{\uparrow} = 0.9$, whereas for top-down modulation $\gamma_{\downarrow} = 30\gamma_{\uparrow}$. Thus, the recognition of the

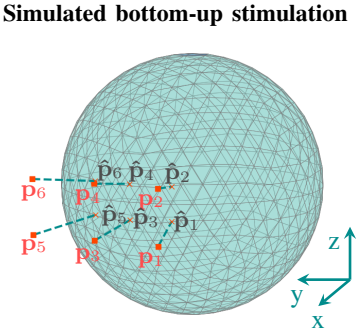

Fig. 3. Bottom-up stimulation at six locations in the sensory ego-space. The objects' center of mass coordinates $\mathbf{p}_i$ and projection $\hat{\mathbf{p}}_i$ on the ego-sphere are shown. To improve visualization, the frame of reference located at the ego-sphere's center is shown at the bottom right.

n^{th} object at time t_w stimulates the model through a unit step function $\lambda = f(t_\text{w}, t_\text{w} + \delta_\text{t})$ with duration $\delta_\text{t} = 1$ sec so

$$\gamma_{n(t)} = \lambda\gamma_\downarrow + (1-\lambda)\gamma_\uparrow. \tag{8}$$

2) Searching around objects: this simulation considered perspective-taking type interactions, where someone indicates a topological reference in another's point of view, such as turning attention to a stimulus in one direction $d \in \{\text{right}, \text{left}, \text{above}, \text{below}\}$, which are words recognized by the robot.

Let $\mathbf{s}_{i(t)}$ in Eq. (1) be modeled such that

$$\mathbf{s}_{i(t)} = \sum_d \gamma_{d(t)} f\left(\gamma_\text{r} m\left(\hat{\boldsymbol{\mu}}_{(t)}, \mathbf{x}_i\right)\right) g\left(\left|\hat{\boldsymbol{\mu}}_{(t)} - \mathbf{x}_i\right|\right) \mathbf{u}_{i(t-1)}. \tag{9}$$

A step function gain $\gamma_{d(t)}$ is set from direction keywords recognition, $f(.)$ is a sigmoid function with gain constant γ_r and $g(.)$ is a multivariate Gaussian function (see Eq. (2)). The mean location reference $\hat{\boldsymbol{\mu}}$ on the ego-sphere representing instantaneous attention selection is obtained so

$$\hat{\boldsymbol{\mu}}_{(t)} = \sum_n \mathbf{k}_{n(t-1)} \hat{\mathbf{p}}_{n(t)}. \tag{10}$$

It is interesting noticing that, by receiving feedback from the output layer $\mathbf{k}_{n(t-1)}$ (see Eq. (6)), local search can be obtained in case the agent is focusing on a region related to a particular object. For horizontal search, the projection's y-coordinate is considered, whereas for vertical search the z-coordinate is more relevant (see Fig. 3). Thus, $m(.)$ in Eq. (9) is defined such that

$$m\left(\hat{\boldsymbol{\mu}}_{(t)}, \mathbf{x}_i\right) = \begin{cases} \hat{\boldsymbol{\mu}}_{\text{y}(t)} - \mathbf{x}_{i\text{y}} : d = \text{left} \\ \mathbf{x}_{i\text{y}} - \hat{\boldsymbol{\mu}}_{\text{y}(t)} : d = \text{right} \\ \hat{\boldsymbol{\mu}}_{\text{z}(t)} - \mathbf{x}_{i\text{z}} : d = \text{below} \\ \mathbf{x}_{i\text{z}} - \hat{\boldsymbol{\mu}}_{\text{z}(t)} : d = \text{above} \end{cases}. \tag{11}$$

3) Losing interest in something: the situation considered here is the agent's loss of interest to an object when receiving negative feedback from the human, after recognizing for instance the word "no". For this, inhibitory activity is induced through a step function gain $\gamma_{q(t)}$ at the pre-selection layer $\mathbf{u}_{(t)}$, receiving feedback from the output layer $\mathbf{k}_{(t-1)}$ (see Eqs. (6), (10)), so the input $\mathbf{s}_{i(t)}$ in Eq. (1) is modeled so

$$\mathbf{s}_{i(t)} = \gamma_{\text{q(t)}} f\left(-g\left(\left|\hat{\boldsymbol{\mu}}_{(t)} - \mathbf{x}_i\right|\right)\right) \mathbf{u}_{i(t-1)}, \tag{12}$$

with $f(.)$ the softmax function and $g(.)$ the multivariate Gaussian function (see Eq. (2)).

B. Experiment

An interaction experiment was designed with the robot Pepper. Landmark stickers recognized by the robot were attached to locations in the scene representing objects (see Fig. 6). The robot was programmed to respond to speech in the vocabulary $V = \{$above, below, left, right, no, one, two, three, four$\}$. The ego-spheres were placed for agents at the center of the torso when in stand-up posture. In this study we do not consider the possibility of rotation of the ego-sphere, which is perhaps acceptable to short face-to-face interactions where participants talk about objects around (see Fig. 1).

The library MediaPipe was used to track the human skeleton from the robot's on-board camera acquisitions. The robot was programmed to keep the human's full body in view during interaction. The human provided feedback to the robot verbally or by pointing to locations in the environment.

During the learning stage, the robot named an object each time and pointed to it for two seconds. The robot's attention model was stimulated as described in the simulation scenario *focusing on named objects*. In order to track the human's non-verbal focus of attention, the intersection of the forearm with the ego-sphere was considered as a cue for object interest. Thereby, the robot learned perspective-taking by associating its tracked version of human's ego-sphere attention selection (modulated by the human's pointing behavior) to its own tracked attention selection state when both are sharing attention to that particular object.

The Hebbian plasticity rule was activated shortly after the robot named the object and both the human and the robot were pointing at it, such that

$$\mathbf{H}_{ij(t)} = \mathbf{H}_{ij(t-1)} + \alpha f\left(\mathbf{a}^{[\text{r}]}_{i(t)}\right) f\left(\mathbf{a}^{[\text{h}]}_{j(t)}\right), \tag{13}$$

where $f(.)$ is the softmax function. The exponent notation in brackets indicates the agent to which the model belongs ([h]uman, [r]obot) with learning rate α. During the interaction stage, since input to the robot's attention model comes partly from the tracked human's attention model, the term $\mathbf{s}^{[\text{r}]}_{i(t)}$ in Eq. (1) is modeled with gain γ_h so

$$\mathbf{s}^{[\text{r}]}_{i(t)} = \gamma_\text{h} \sum_j \mathbf{H}_{ij(t)} \text{softmax}\left(\mathbf{a}^{[\text{h}]}_{j(t)}\right). \tag{14}$$

Notice that inputs from bottom-up saliency and top-down modulation sources are added together.

C. Materials and Resources

The hardware components included a computer with 64 GB RAM memory, 11$^{\text{th}}$ Generation Intel® Core™ i9-11950H @ 2.60GHz × 16, and graphic card NVIDIA RTX A4000. The project counted on a humanoid robot Pepper, manufactured by Softbank Robotics. The software components were implemented in Python programming language versions 2.7 and 3, running in Ubuntu (20.04 LTS). The library MediaPipe version 0.10.3 was used to track the human posture from monocular vision. The library *naoqi* version 2.5.7.1 was employed for the control programs.

V. RESULTS

Globally, the studies[2] in simulation showed the possibility of representing bottom-up saliency and top-down attention modulation at the pre-selection layer. Figure 4 illustrates how considering feedback from the focus-to-object output layer is a convenient means for inducing neural dynamics at particular regions in the ego-sphere, so helping the agent to respond to embodied references in verbal communication.

Figure 5 presents 60 seconds of the simulated interaction. Here, the keywords *left*, *right*, *above* and *below* were employed so the agent could focus on the six stimuli one at a time. Similar results were obtained for the other two scenarios: *focusing on named objects* and *losing interest in something*. It is interesting noticing on the plot at the center how attention selection is obtained through a competition process, resulting from inhibitory synaptic interaction.

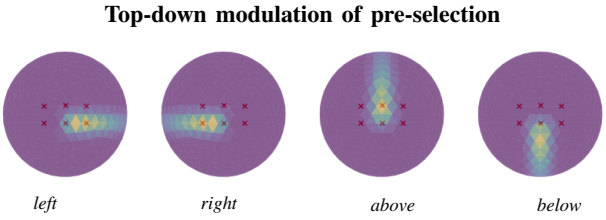

Fig. 4. Simulation scenario *searching around objects*. Top-down modulation of attention at the pre-selection layer is shown after instantaneous recognition of the keywords *left*, *right*, *above*, and *below* see (Eq. (9)), relative to the location $\hat{\mathbf{p}}_3$ in Fig. 3).

Concerning the experiment (see Fig. 6), results showed that during the learning stage the human was able to learn the objects' name from the robot, whereas the robot was able to learn the body relation of the human to such objects. A total of seven trials were recorded. It is important to mention that the human was only instructed to stand between the landmarks while facing the robot. No marks were attached to the floor to avoid rigorously determining the human's position, thus introducing some variability in trials. Figure 7 presents the comparison of attention modulation from learning only in the first trial (see Eq. (13)), so object saliency is induced on other trials without re-enabling learning (see Eq. (14)). Thus, it can be noticed that the Hebbian plasticity rule was able to generalize for other trials in the experiment.

[2]Video available at: `https://youtu.be/zjzpy7lpZlg`

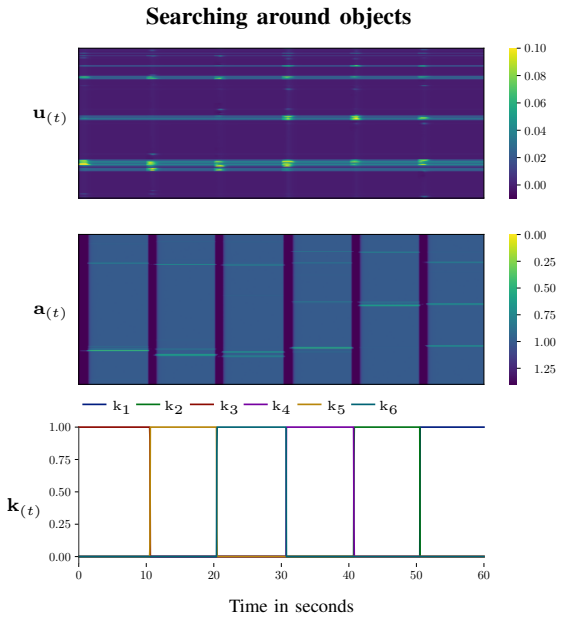

Fig. 5. Focusing initially on object 3 (relative to $\hat{\mathbf{p}}_3$, see Fig. 3), the agent switched attention to surrounding objects, from instantaneous recognition of keywords in the sequence: *right* ($t = 10$), *above* ($t = 20$), *left* ($t = 30$), *left* ($t = 40$) and *below* ($t = 50$). For more details, see Eqs. (9)(11).

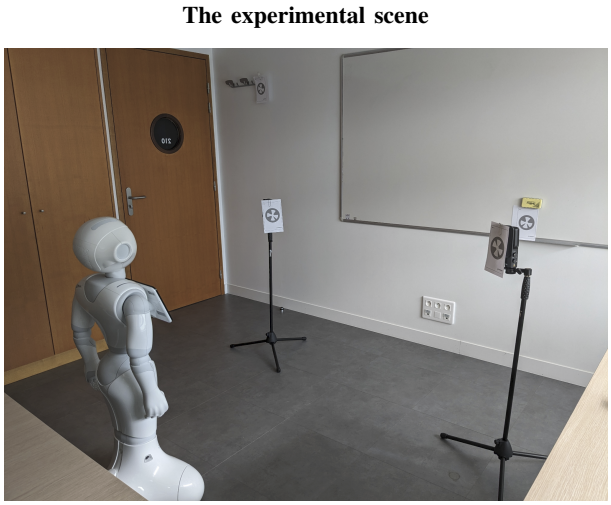

Fig. 6. Landmarks recognized by Pepper were set up in the room to mimic objects. The human stood in front of the robot.

Figure 8 shows a sequence of pointing gestures captured from the robot's on-board camera. The robot tracked both ego-sphere online during the experience. In the situations depicted, the human's pointing behavior was able to modulate attention at a pre-selection level on the robot (see Eq. (14)), relative to stimuli salient in the environment.

VI. CONCLUSIONS

This work focused on a critical gap in social robotics and HRI which are current limitations of robots to engage in intuitive forms of interaction with humans and adapt to

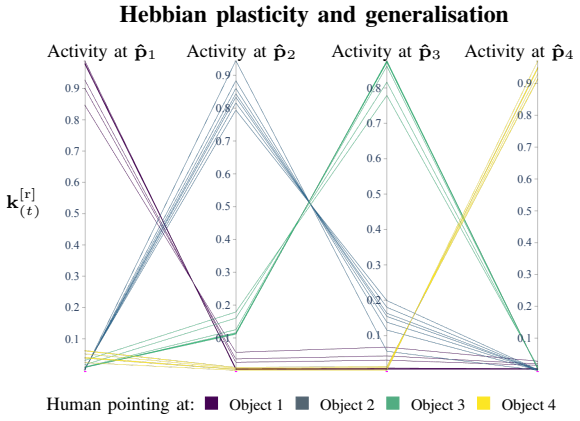

Fig. 7. Instantaneous effect of Hebbian learning on the robot's focus-to-object layer (see Eq. 6), as detailed in Eq. (14) ($\gamma_{\mathrm{h}} = 2.5\mathrm{e}3$). Despite variability in positioning, after learning a single trial, the human's pointing gesture can clue the robot's attention in other trials (seven in total).

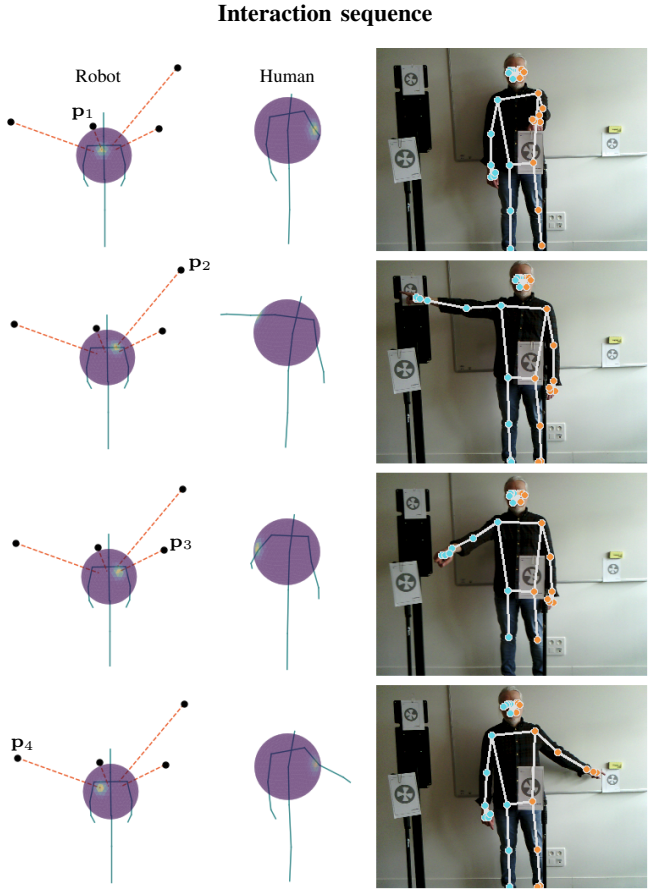

Fig. 8. Landmarks' position (black dots) and projection ray (red) are shown in the robot's ego-sphere. The human is tracked from the robot's camera (right) at about 10 Hz. After the learning stage, the human's pointing gesture is able to stimulate locations in the robot's ego-sphere which are proximal to landmarks' projection center. This helps the robot to focus on such objects.

situations encountered in everyday life. Consequently, we targeted the core aspect of communication skills required for robots to share attention and knowledge with humans.

Given our interest in studying attention selection dynamics from low-level cognitive processes, we explored in the bio-inspired literature the concept of *ego-sphere*. After carefully considering previous works, we summarized limitations as follows: a) not investigating sufficiently interaction dynamics between locations in the ego-sphere, b) being an approach too egocentric for HRI by considering the robot as the only agent provided with an ego-sphere, and c) with very few exceptions, neglecting the aspect of compositionality in knowledge representation. Therefore, we proposed AEGO as a three-layer neural architecture to track attention selection for each agent, inspired by 4E cognition, FIT, and DNF research. We showed how bottom-up saliency, top-down generative modulation, and lateral connectivity interact to obtain online attention selection.

We believe that AEGO helps advance the field of HRI in two important ways. Firstly, by carefully considering the implications of Moravec's Paradox, we showed how inspiration on 4E cognition principles resulted in communication skills prototypes relying on low-level sub-symbolic representations, thus avoiding the high cost of acquiring extensive knowledge about the environment, which is characteristic of disembodied symbolic approaches. Also, since the model is differentiable, it can benefit from well-established machine learning methods to complement or improve existing neural network models, for tasks requiring attention tracking.

Secondly, by proposing a common representation space for individual attention and describing how information can be fused to account for the influence of others' behavior on attention selection as a generative modulation process, AEGO can contribute to adaptation to HRI instantaneous emerging relations by allowing the robot not only to share attention to objects, but potentially detecting variations on attention sharing (or *surprise*, as described in free-energy principle theory [11]), resulting from intention changes in the human. This is a fundamental competence for robot agency that opens interesting perspectives to investigate forms of intersubjectivity (see [5]) mediated by JA.

Our study presented some limitations that must be addressed. In future research, AEGO should be studied within group interaction situations where attention is shared under the influence of social affordance and personal style. Thus, more complex forms of JA modulation unfolding within social contexts should be considered. We believe that integrating AEGO and TOP-JAM [7] would be a relevant step in this direction. Another perspective is exploring how attention sharing dynamics can help engagement and agency from the initiation, response or maintenance of JA in HRI. Also, since objects were static in the environment, it should be investigated how moving objects or rhythmic interaction can influence attention sharing.

ACKNOWLEDGMENT

We thank the IDMC (University of Lorraine) for providing us with the robot for the experiments and the ANITI Project for funding the early stage of this collaboration.

REFERENCES

[1] ALBUS, J. S. Outline for a theory of intelligence. *IEEE transactions on systems, man, and cybernetics 21*, 3 (1991), 473–509.

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

[3] BELHASSEIN, K., FERNÁNDEZ-CASTRO, V., MAYIMA, A., CLODIC, A., PACHERIE, E., GUIDETTI, M., ALAMI, R., AND COCHET, H. Addressing joint action challenges in hri: Insights from psychology and philosophy. *Acta Psychologica 222* (2022), 103476.

[4] BODIROZA, S., SCHILLACI, G., AND HAFNER, V. V. Robot ego-sphere: An approach for saliency detection and attention manipulation in humanoid robots for intuitive interaction. In *2011 11th IEEE-RAS International Conference on Humanoid Robots* (2011), IEEE, pp. 689–694.

[5] 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.

[6] CHAME, H. F., AND CHEVALLEREAU, C. Grounding humanoid visually guided walking: From action-independent to action-oriented knowledge. *Information Sciences 352* (2016), 79–97.

[7] CHAME, H. F., CLODIC, A., AND ALAMI, R. Top-jam: A bio-inspired topology-based model of joint attention for human-robot interaction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (2023).

[8] 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.

[9] DOMHOF, J., CHANDARR, A., RUDINAC, M., AND JONKER, P. Multimodal joint visual attention model for natural human-robot interaction in domestic environments. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2015), IEEE, pp. 2406–2412.

[10] ELDARDEER, O., GONZALEZ-BILLANDON, J., GRASSE, L., TATA, M., AND REA, F. A biological inspired cognitive framework for memory-based multi-sensory joint attention in human-robot interactive tasks. *Frontiers in Neurorobotics 15* (2021), 648595.

[11] FRISTON, K., SCHWARTENBECK, P., FITZGERALD, T., MOUTOUSSIS, M., BEHRENS, T., AND DOLAN, R. J. The anatomy of choice: active inference and agency. *Frontiers in human neuroscience 7* (2013), 598.

[12] GROTZ, M., HABRA, T., RONSSE, R., AND ASFOUR, T. Autonomous view selection and gaze stabilization for humanoid robots. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2017), IEEE, pp. 1427–1434.

[13] HEIKKILÄ, P., LAMMI, H., NIEMELÄ, M., BELHASSEIN, K., SARTHOU, G., TAMMELA, A., CLODIC, A., AND ALAMI, R. Should a robot guide like a human? a qualitative four-phase study of a shopping mall robot. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11* (2019), Springer, pp. 548–557.

[14] ITTI, L., KOCH, C., AND NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*, 11 (1998), 1254–1259.

[15] KOCH, C., AND ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology 4*, 4 (1985), 219–227.

[16] MARQUES-VILLARROYA, S., CASTILLO, J. C., GAMBOA-MONTERO, J. J., SEVILLA-SALCEDO, J., AND SALICHS, M. A. A bio-inspired endogenous attention-based architecture for a social robot. *Sensors 22*, 14 (2022), 5248.

[17] NEWEN, A., GALLAGHER, S., AND DE BRUIN, L. 34E Cognition: Historical Roots, Key Concepts, and Central Issues. In *The Oxford Handbook of 4E Cognition*. Oxford University Press, 09 2018.

[18] PETERS, R. A., HAMBUCHEN, K. A., AND BODENHEIMER, R. E. The sensory ego-sphere: a mediating interface between sensors and cognition. *Autonomous Robots 26*, 1 (2009), 1–19.

[19] RUESCH, J., LOPES, M., BERNARDINO, A., HORNSTEIN, J., SANTOS-VICTOR, J., AND PFEIFER, R. Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *2008 IEEE International Conference on Robotics and Automation* (2008), IEEE, pp. 962–967.

[20] SALLAMI, Y., LEMAIGNAN, S., CLODIC, A., AND ALAMI, R. Simulation-based physics reasoning for consistent scene estimation in an hri context. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2019), IEEE, pp. 7834–7841.

[21] 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.

[22] SCHÖNER, G., AND SPENCER, J. P. *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press, 2016.

[23] SERHAN, B., SPENCER, J., AND CANGELOSI, A. Coupling dynamical and connectionist models: Representation of spatialattention via learned deictic gestures in human-robot interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2018), vol. 40.

[24] SIAGIAN, C., CHANG, C. K., AND ITTI, L. Autonomous mobile robot localization and navigation using a hierarchical map representation primarily guided by vision. *Journal of Field Robotics 31*, 3 (2014), 408–440.

[25] SIPOSOVA, B., AND CARPENTER, M. A new look at joint attention and common knowledge. *Cognition 189* (2019), 260–274.

[26] TREISMAN, A. M., AND GELADE, G. A feature-integration theory of attention. *Cognitive psychology 12*, 1 (1980), 97–136.