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

Hendry F. Chame[1] and Rachid Alami[2]

***Abstract*— Despite important progress in recent years, social robots are still far away from showing sophisticated skills of interaction and adaptation to human environments. Our research is concerned with the study of social cognition in HRI, notably with communication skills relying on joint attention (JA) and knowledge sharing. Since JA involves low-level cognitive skills, we take into account the implications addressed by Moravec's Paradox and focus on the aspect of knowledge representation. Inspired by embodiment and 4E cognition principles, we study egocentric localization through the concept of sensory *ego-sphere*. We propose a neural network model of attention selection named AEGO from *dynamic fields theory*, which takes into account the dynamics of bottom-up and top-down modulation processes and the effects of neural excitatory and inhibitory synaptic interaction. We studied the model in simulations and analyzed some application scenarios in HRI. We then conducted a real experiment for a JA-based task with the robot Pepper, considering proprioception, vision and basic natural language recognition. Results showed that AEGO is a convenient representation for HRI allowing the human and the robot to share attention and knowledge about objects in the environment.**

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 a one-year-old child. Behind this paradox remains the question in artificial intelligence research of what sort of knowledge representation would be suitable for allowing 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 environment in interaction; a perspective known as *4E cognition* [11].

Inspired by embodiment and 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 skill 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 [17].

[1]LORIA-CNRS (NeuroRhythms team). Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France. hendry.ferreira-chame@loria.fr.

[2]LAAS-CNRS (Robotics and InteractionS team). 7 Av. du Colonel Roche, 31400 Toulouse, France. rachid.alami@laas.fr.

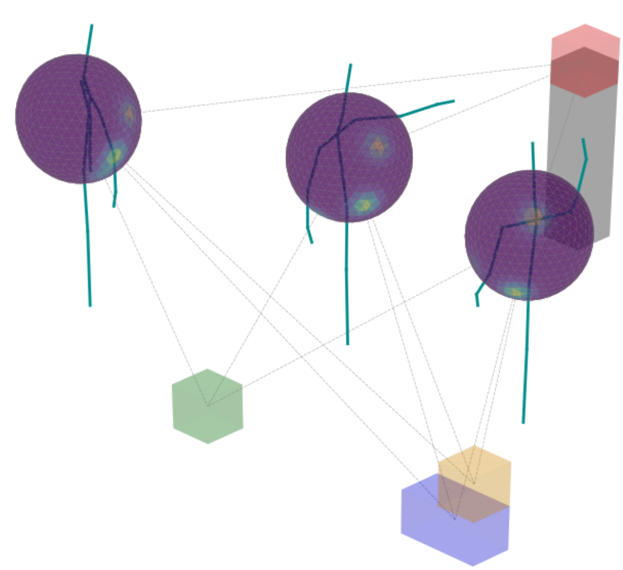

Fig. 1. The scene shows a situation where two agents are interacting about the orange object, whereas another agent is leaving the scene. Three other objects are present the effect of each object agents ego-spherical localization at a pre-selective stage is shown.

As a continuation of previous research in which we proposed a model for tracking JA in HRI as a topology-based representation organized as a *scale of jointness* [5], in this work we investigate the more fundamental aspect of attention selection and how such mechanism could allow the emergence of JA in human-robot interaction from top-down and bottom-up modulation processes. For this, as shown in Fig. 1, we suppose unconstrained situations where agents can become interested in objects in the environment and eventually share attention and knowledge about them (e.g. situations like asking someone for direction or commenting about salient stimuli like a noise or an object).

Tacking into account the considerations above, we explore the concept of *ego-sphere* [1] and propose the model named AEGO for tracking the attention focus of agents as represented from egocentric perspective, and resulting from on-board sensory acquisitions. For this, inspired by *dynamic neural fields* (DNF) theory [2], we model attention selection as dynamical system process represented by neural fields networks with lateral connectivity. By addressing limitations on previous research, we show how neural excitatory and inhibitory interaction allows 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 an interesting resource for HRI applications.

This document is organized as follows: Section II discusses previous works and argues 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 tracking from top-down and bottom-up modulation processes, and showing potential applications, and b) conducting an experiment with the robot Pepper including proprioception, vision and basic natural language recognition about relations in the environment. Section V reports on the study's results, and Section VI presents conclusions and future perspectives.

II. PREVIOUS WORK

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 [13] 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 [3] focused on intuitive HRI, including the possibility of top-down modulation of attention. The aspect of information representation has also been studied in [12], so the ego-sphere has been implemented as a storage data-base indexed by spherical tessellation mapping. Other contributions could be mentioned (e.g. [6], [10]).

To our knowledge, previous research has not explored sufficiently the aspect of interaction dynamics between locations represented in the ego-sphere, considering at most basic forms of interaction spread between nodes. Moreover, excluding saliency map approaches (e.g. [13]), the dynamics of attention was 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 research is considering the robot as the only one in interaction given with embodied sensory ego-spherical representations, so data coming from human agents is mostly represented in the robot's perspective. In our opinion, this would be a too egocentric approach for HRI. We believe that when keeping track of embodied relations between agents and objects in the environment as a distributed dynamical system, the robot could take decisions and behave without relying too much on environment modeling, from the emergence of attention from instantaneous interaction. Hence, we propose that attention selection is tracked simultaneously from participants' egocentric perspective.

Our previous research has also constituted relevant steps in the direction of developing the current study which is worth mentioning. In [4] an ego-cylindrical selection mechanism for attention was proposed for autonomous positioning with respect to objects in the environment. In [5] the model TOP-JAM was proposed as a means for JA tracking in HRI from allocentric references. In [7] joint attention in HRI is studied for a providing guidance task in a shopping mall. [Many papers from the RIS team could be cited: human-aware planning, JA ... To be confirmed with Rachid].

To summarize, we propose to model attention for HRI in neural dynamic fields networks for tracking the influence of three important sources on attention selection: a) bottom-up stimulation, b) top-down modulation, and c) local interactions from inhibitory and excitatory synapse. We named this network AEGO and show how it is a useful representation for tracking attention in HRI, which can be conveniently included in several experimental setups. We present in the section below the mathematical foundations of the model. In Section IV we show how AEGO is suited for investigating joint attention in HRI.

III. THE MATHEMATICAL MODEL

Theoretical models of visual attention such as *feature integration theory* (FIT) [18], have described attention as a multi-level information fusion process. According to FIT, at a pre-attention level the perceptual system receives from separate maps encoding feature salience (e.g., color, edges, shapes), which are lately combined at an attention selection stage. A bio-inspired architecture has been proposed from FIT by [9] and implemented by [8], finding applications in robotics (e.g. autonomous navigation based on vision [16]).

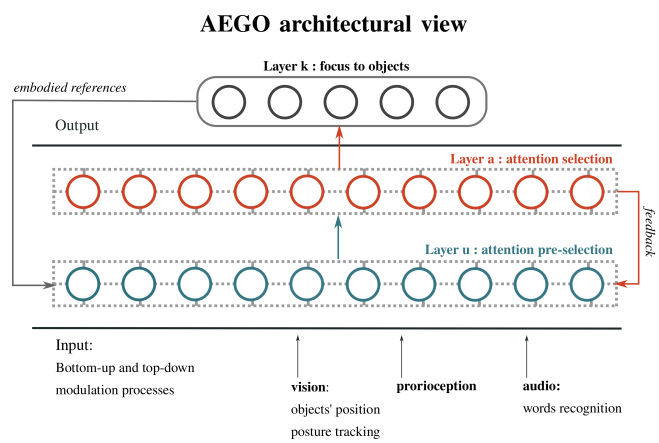

Fig. 2. The AEGO model, as detailed in Eqs. (1), (3) and (6).

Under the influence of previous research, we propose a model of attention selection inspired by FIT and DNF theory. [We focus not only in vision but on the agent's capacity of recognizing stimuli and roughly calculating their relation to the body in ego-spherical representations]. Thus, we consider a pre-attention phase where proprioceptive and exteroceptive stimuli excite the ego-space topology, encoded by dynamic fields neural networks, receiving inputs from top-down and bottom-up processes and synaptic interaction. In a posterior stage, attention selection results from competition in synaptic interaction. Hence, we propose the model named AEGO which architectural view is illustrated in Fig. 2. The mathematical definition of the model [1] is detailed next.

A. Pre-attention phase

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

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

According to the principle of local interconnections [14], the interaction strength $\mathbf{U}_{ij}$ between neurons i and j is selected so proximal locations have stronger interaction. Hence, multivariate Gaussian weights are selected, such that

$$\mathbf{U}_{ij}(|\mathbf{x}_i - \mathbf{x}_j|) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{t}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right)}{\sqrt{2\pi^3|\boldsymbol{\Sigma}|}} \quad (2)$$

The term $\mathbf{s}_{i(t,\Xi)}$ represents the stimulation received at time instant t affecting the ego-space locations Ξ. As it is going to be detailed in Section IV, this term models stimulation from both top-down and bottom-up processes. Finally, in Eq. (1) q_{u} corresponds to the activation resting state, τ_{u} is the a time constant, and ϵ is a global inhibition factor affecting lateral interactions between neurons.

B. Attention selection phase

Let the activation of the i^{th} neuron represent the dynamics of an attention selection process at a particular location in the ego-space, such that

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

Inhibitory neural interaction has been associated with selection mechanisms [15]. Thus, we propose to model lateral interaction $\mathbf{A}_{ij}$ between neuron i and j such that

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

with $\varphi = \max(\mathbf{U}_{i:})^{-1}$ a scaling factor. The activation function f in Eq. (3) is defined so

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

being γ_{u} and α gain constants.

[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 columns vector (e.g. $\mathbf{A}_{i:}$ for columns and $\mathbf{A}_{:i}$ for rows). Position and orientation vectors are in 3D Cartesian space unless stated otherwise. The projection of a 3D position $\mathbf{p}$ in the ego-sphere surface is denoted $\hat{\mathbf{p}}$. The dot product between vectors $\mathbf{p}$ and $\mathbf{v}$ is denoted $\mathbf{p} \cdot \mathbf{v}$.

C. Object focus output layer

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

$$\mathbf{k}_{\mathrm{i(t)}} = \mathrm{softmax}\left(\gamma_{\mathrm{k}}\sum_j \mathbf{O}_{ij(t)}\mathbf{a}_{j(t)}\right) \quad (6)$$

where $\mathbf{O}_{ij}(|\hat{\mathbf{p}}_i - \mathbf{x}_j|)$ is obtained from Eq. (2) with $\hat{\mathbf{p}}_i$ the projection of the object's center of mass in the ego-sphere, and γ_{k} represents a gain factor constant.

IV. METHODOLOGY

We designed several studies in simulation for testing AEGO and analyzing potential application scenarios. We also conducted a real experiment with the robot Pepper considering propioception, vision and basic natural language recognition. The details of the methodology are next.

A. Materials and Resources

The hardware components included a computer with 64 GB RAM memory, 11^{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). 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 for tracking the human skeleton from monocular vision. The library *naoqi* version 2.5.7.1 was employed for implementing the control programs for Pepper and the software Choregraphe version 2.8.6.23 was used for simulations.

B. Simulations

Table I presents common parameters for the pre-selection, selection and output layers described in Eqs. (1), (3) and (6). The state of the network is obtained by numerical integration by the Euler method, according to the time-step dt. As shown in Fig 3, six objects where simulated as bottom-up sources of stimulation to the agent. Three interactions situation relying of the recognition of basic words where considered to explore top-down modulation of attention. These situations are described below.

1) Focus on a specific object: we investigated the possibility of attending to a specific object as modulated by top-down processes. For this, it is assumed that the agent is able to track and recognize objects in the scene while associating unique words for addressing to them. A numerical ID could serve this purpose. Thus, once the human says "three", attention should be directed to location $\hat{\mathbf{p}}_3$ as shown in Fig. 3. For this, the term $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) can be set so

$$\mathbf{s}_{i(t,\Xi)} = \sum_j \gamma_{\mathrm{o}j}\mathbf{O}_{ij(t)}\left(|\hat{\mathbf{p}}_{j(t)} - \mathbf{x}_{i(t)}|\right) \quad (7)$$

Interest to the j^{th} object is modeled through the gain $\gamma_{\mathrm{o}j}$. For bottom-up saliency $\gamma_{\mathrm{bu}} = 0.9$ for all detected objects,

TABLE I

COMMON PARAMETERS FOR SIMULATIONS

Parameter	Value
Ego-space vertex number	642
Ego-space faces number	1280
dt	50 ms
ϵ	0.0001
$\tau_{\mathrm{u}}, \tau_{\mathrm{a}}$	200 ms
$\mathbf{\Sigma}$	$0.01\mathbf{I}_3$
q_{u}	-0.01
qa	-0.0001
γ_{k}	250
γ_{u}	2.5
γ_{n}	15.5
α	100

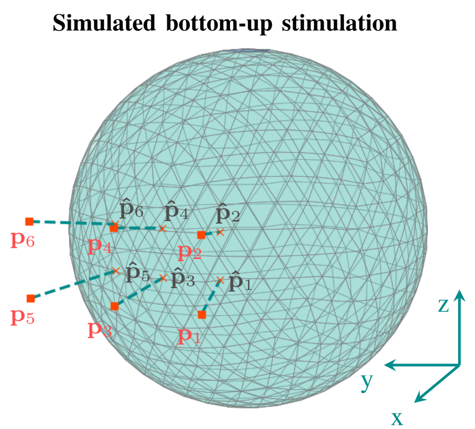

Fig. 3. Bottom-up stimulation at six locations in the sensory ego-space, The objects' center coordinates $\mathbf{p}_i$ and projection $\hat{\mathbf{p}}_i$ on the sensory ego-space are shown. To improve visualization, the frame of reference is shown at bottom-right but it is located at the ego-sphere's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

whereas for top-down modulation it is set to $\gamma_{\mathrm{td}} = 12\gamma_{\mathrm{bu}}$. As the object's name is recognized with probability w, when exceeding a given threshold it influences the model according to a unit step function $\lambda = f(t_{\mathrm{w}}, t_{\mathrm{w}} + \delta_{\mathrm{t}})$, thus

$$\gamma_{\mathrm{o}j} = \lambda\gamma_{\mathrm{td}} + (1-\lambda)\gamma_{\mathrm{bu}} \tag{8}$$

The duration $\delta_{\mathrm{t}} = 1$ sec. The local influence of stimuli in the neural field $\mathbf{O}_{ij(t)}$ is set conforming to Eq. (2).

2) Searching around and object: this simulation considered interactions base on perspective-taking, where someone indicates a topological reference in agent's perspective such as turning attention to a stimulus at *right*, *left*, *above* or *below*, which are terms recognized by the robot. Thus, $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) is modeled such that

$$\mathbf{s}_{i(t,\Xi)} = \sum_j f\left(\gamma_{\mathrm{r}} m_{(t)}\right)\ g\left(\left|\hat{\boldsymbol{\mu}}_{(t)} - \mathbf{x}_{i(t)}\right|\right) \tag{9}$$

where $f(.)$ is the sigmoid function with γ_{r} representing a gain constant and $g(.)$ is the multivariate Gaussian function (see Eq. (2)). The coordinates of the projection $\hat{\boldsymbol{\mu}}$ on the ego-sphere, representing attention selection, are obtained so

$$\hat{\boldsymbol{\mu}}_{(t)} = \sum_i \mathbf{k}_{\mathrm{i(t-1)}}\hat{\mathbf{p}}_{i(t)} \tag{10}$$

It is interest noticing that by considering feedback from the output layer $\mathbf{k}_{\mathrm{i(t-1)}}$ at previous time step (see Eq. (6)), a local search could be achieved if the agent was actually focusing on a particular object. For the case of horizontal search, the y coordinate of the points projection are considered, whereas for vertical search the z coordinate would be of more relevance. Therefore, $m_{(t)}$ in Eq. (9) is selected so

$$m_{(t)} = \begin{cases} \hat{\boldsymbol{\mu}}_{\mathrm{y}(t)} - \hat{\mathbf{p}}_{j\mathrm{y}(t)} : \text{“right”} \\ \hat{\mathbf{p}}_{j\mathrm{y}(t)} - \hat{\boldsymbol{\mu}}_{\mathrm{y}(t)} : \text{“left”} \\ \hat{\boldsymbol{\mu}}_{\mathrm{z}(t)} - \hat{\mathbf{p}}_{j\mathrm{z}(t)} : \text{“above”} \\ \hat{\mathbf{p}}_{j\mathrm{z}(t)} - \hat{\boldsymbol{\mu}}_{\mathrm{z}(t)} : \text{“below”} \end{cases} \tag{11}$$

Similarly to previous scenario, the gain $\gamma_{\mathrm{o}j}$ is set conforming to Eq. (8) with duration $\delta_{\mathrm{t}} = 1$ sec.

3) Focus on another object: the situation considered here is the agent's lost of interest to an object form receiving negative feedback from the human. For this, inhibitory feedback from the selection layer $\mathbf{a}_{(t-1)}$ is provided to the pre-selection later $\mathbf{u}_{(t)}$ (see Eqs. (1),(3)). The term $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) is modeled with a gain constant γ_{n}, such that

$$\mathbf{s}_{i(t,\Xi)} = -\mathrm{softmax}\left(\gamma_{\mathrm{n}}\mathbf{a}_{i(t-1)}\right) \tag{12}$$

The gain $\gamma_{\mathrm{o}j}$ is set conforming to Eq. (8) but the duration of stimulation is selected shorter for this case $\delta_{\mathrm{t}} = 0.5$ sec.

C. Experiment

An interaction experiment was designed with the robot Pepper. Since the robot is capable of recognizing typical landmarks, some were attached to locations on the environment representing objects. The robot is also capable of speech recognition within dialog context, so it was programmed to recognize the terms *above*, *below*, *left*, *right*, *no*, *one*, *two*, *three*, and so on. The ego-space was placed at the robot's *Torso* frame, when in Stand-up posture. In this study we do not consider the possibility of rotation and translation of the ego-sphere (which would require of geometrical remapping), so the experiment would correspond to situations of short interactions where participants talk about objects around.

[I got until here ...]

V. RESULTS

VI. CONCLUSIONS

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).

Searching around operators

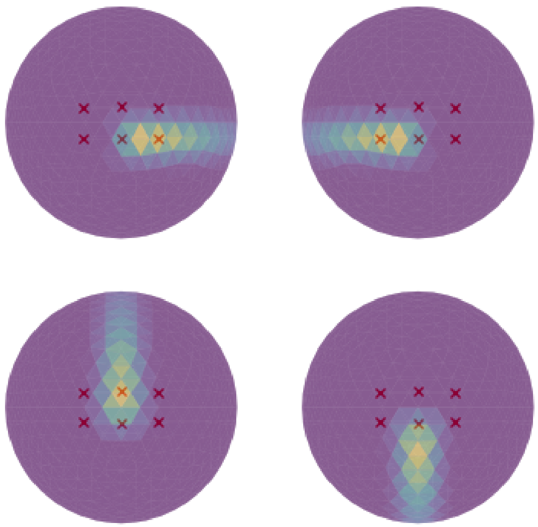

Fig. 4. As described in the simulation *searching around an object*, the top-down modulation of attention is shown after instantaneous recognition of the words *left*, *right*, *above*, and *below* see (Eq. (9)), relative to $\hat{\mathbf{p}}_3$.

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] 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.

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

[5] 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).

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

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