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

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 adaptation, interaction and communication 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. By embracing embodiment and 4E cognition principles, we study egocentric localization through the concept of *sensory ego-sphere*. Inspired by *dynamic fields theory*, we propose a neural network model of attention selection named AEGO, which takes into account the dynamics of bottom-up and top-down modulation processes and the effects of neural exitatory and inhibitory synaptic interaction on attention selection. We studied the selection model in simulations and analyzed some application scenarios in HRI. We then conducted a real experiment of 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].

In agreement with the perspective of 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 such processes be modulated in direct interaction. Moreover, we study JA as a multi-dimensional construct involving cognitive

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

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

skills so constituting forms of social attention at distinct levels of interaction and knowledge sharing [17].
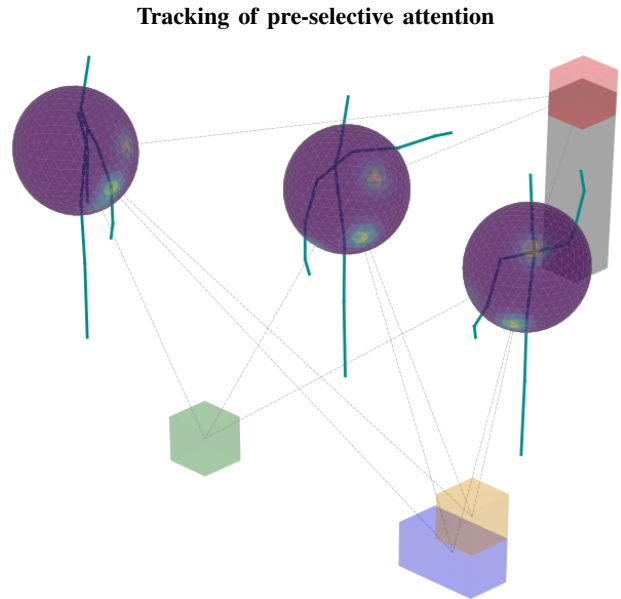
Tracking of pre-selective attention

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 within a topology-based representation constituting 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 on the environment and eventually share attention and knowledge about it (e.g. situations like asking someone for direction or commenting about salient stimulus 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 acquisitions of robot's on-board sensory. For this, inspired on *dynamic neural fields* (DNF) theory [2], we model attention selection as a dynamical system process represented by a neural network with lateral conectivity. By addressing limitations on previous research, we show how

neural exitatory and inhibitory interaction allows to study the emergence of attention selection. Moreover, we show how the model can be used to track agents' interaction with peripersonal space, which becomes an interesting resource for HRI-based 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 of the work, consisted in: a) studying in simulation the attention tracking mechanism proposed, which relies on the selection process, and showing potential applications, and b) conducting an experiment with the robot Pepper including propioception, vision and basic natural language recognition about relations in the environment. Section V reports the study's results, and Section VI presents conclusions and future perspectives.

II. PREVIOUS WORK

According to [1] an *ego-sphere* consists in 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 showed how attention and short-term memory can be modulated through saliency maps and allow the robot to explore the environment exploration based on novelty. A work by [3] focused on intuitive HRI, including the possibility of top-down modulation of attention. The aspect of 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 based on these ideas could be mentioned (eg. [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, excusing 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 ego-sensory 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 would be able to take more informed decisions for operating in such environment. 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 object 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

Feature integration theory (FIT) [18] of visual attention studies the role attention plays in selecting and fusing complex information. According to FIT, at a pre-attention level the perceptual system is constituted by separate maps, each encoding feature salience (e.g., color, edges, shapes) which are lately combined at a attention stage. A biologically plausible architecture has been proposed from FIT in [9], an the implementation by [8] has been employed for visually-guided autonomous navigation (e.g. [16]).
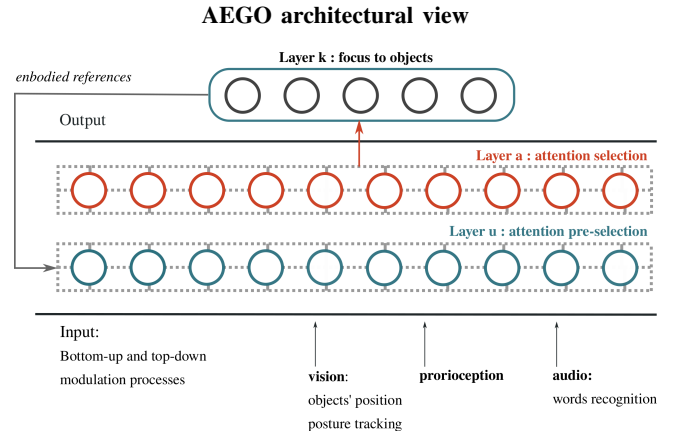

Fig. 2.   The AEGO model, as detailed in Eqs. (1), (3) and (6).

Let the agent's peripersonal space be represented by a localization topology defined by a vertex set, resulting from the tessellation of an icosahedron polyhedron, which approximates a spherical region around the agent. Under the influence of previous research, we propose a model of attention selection inspired by FIT and DNF theory [2]. For this we consider a pre-attention phase where proprioceptive and exteroceptive stimuli excite the ego-space topology, encoded by dynamic field neural network, receiving inputs from top-down and bottom-up processes and synaptic interaction. In a posterior stage, attention selection results from a competition

process ruled by inhibitory synaptic interaction. The neural network structure is illustrated in Fig. 2 and the mathematical model [1] is detailed below.

A. Pre-attention phase

Let the activation of the ith 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

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

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

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

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 (1) h_{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 ith neuron represent the dynamics of an attention selection process at a particular location in the ego-space, such that

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

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 (3) is defined so

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

being ς and α gain constants.

---

[1] **Notation.** Matrices and vectors are represented in bold, indexes are represented as subscripts (e.g. the ith 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 ith 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 of attending to particular objects $\mathbf{k}_{\text{i}(t)}$ be modeled as an output layer, such that

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

where 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 proprioception, 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$^{\text{th}}$ 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). For the simulated scenarios, the state of the network is obtained by numerical integration with the Euler method, according to the time-step dt. As shown in Fig 3 six objects where simulated as sources of stimulation to the agent. In total, three interactions considering the recognition of basic words where studied, which is described next.

TABLE I

COMMON PARAMETERS FOR SIMULATIONS

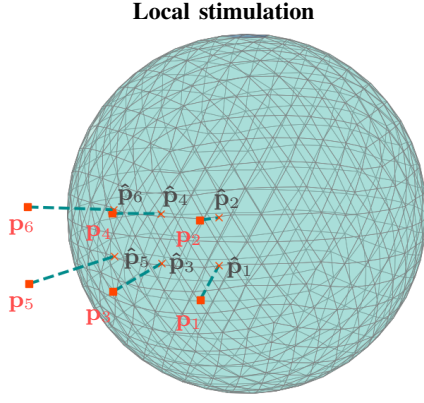| Parameter	Value
Ego-space vertex number	642
Ego-space faces number	1280
dt	50 ms
ϵ	0.0001
$\tau_{\text{u}}, \tau_{\text{a}}$	200 ms
$\boldsymbol{\Sigma}$	$0.01\mathbf{I}_3$
h_{u}	-0.01
h_{a}	-0.0001
γ_{k}	250
α	100
ς	2.5

Local stimulation

Fig. 3. Stimulation was simulated at six locations in the sensory ego-space. Objects' center coordinates $\mathbf{p}_i$ and corresponding projection $\hat{\mathbf{p}}_i$ on the sensory ego-space are shown.

1) Focus on a specific object: we investigated the possibility of attending to a specific object as modulated by top-down processes. For this, the term $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) was modeled such that

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

with $\hat{\mathbf{p}}_i$ the projection on the ego-space of objects' estimated center of mass coordinates. Interest was modeled through the gain γ_j. For bottom-up saliency it was set to $\gamma_{bu} = 0.9$, whereas for top-down modulation of interest it was set to $\gamma_{td} = 12\gamma_{bu}$, acting on the model through a one second step function. The local influence of stimulus $\mathbf{O}_{ij}$ is selected conforming to Eq. (2).

2) Search around and object: this simulation considered an interaction situation based on proxemics where someone takes the agent's perspective and indicates a topological reference such as turning attention to the another stimulus located at right, left, above or below in egocentric perspective. Hence, the term $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) was modeled so

$$\mathbf{s}_{i(t,\Xi)} = \sum_j f\left(\gamma_r m_{(t)}\right) \; g\left(\left| \hat{\boldsymbol{\mu}}_{(t)} - \mathbf{x}_{i(t)} \right| \right) \quad (8)$$

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 is defined so

$$\hat{\boldsymbol{\mu}}_{(t)} = \sum_i \mathbf{k}_{i(t-1)} \hat{\mathbf{p}}_{i(t)} \quad (9)$$

In horizontal search, the y coordinate of the points are considered, whereas in vertical search the z coordinate is more relevant. Therefore, $m_{(t)}$ in Eq. (8) is selected so

$$m_{(t)} = \begin{cases} \hat{\boldsymbol{\mu}}_{y(t)} - \hat{\mathbf{p}}_{jy(t)} : \text{right} \\ \hat{\mathbf{p}}_{jy(t)} - \hat{\boldsymbol{\mu}}_{y(t)} : \text{left} \\ \hat{\boldsymbol{\mu}}_{z(t)} - \hat{\mathbf{p}}_{jz(t)} : \text{above} \\ \hat{\mathbf{p}}_{jz(t)} - \hat{\boldsymbol{\mu}}_{z(t)} : \text{below} \end{cases} \quad (10)$$

3) Focus on another object: in this simulation scenario the robot loses attention to an object form receiving negative feedback from the human. Thus, let the term $\mathbf{s}_{i(t,\Xi)}$ in Eq. (1) be modeled considering the gain constant γ_w such that

$$\mathbf{s}_{i(t,\Xi)} = -\text{softmax}\left(\gamma_w \mathbf{a}_{i(t)}\right) \quad (11)$$

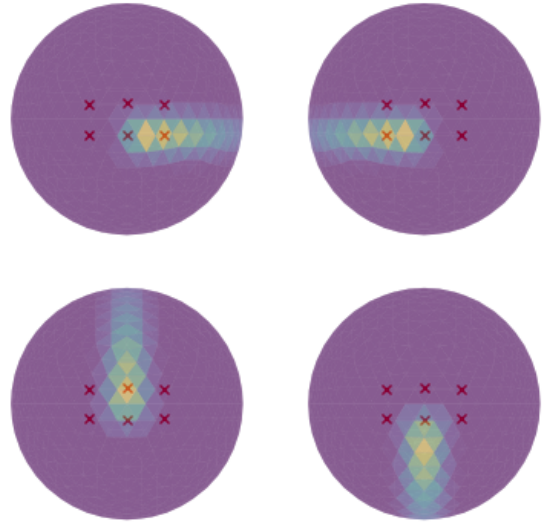*C. Experiment*

V. RESULTS

Searching around operators

Fig. 4. The AEGO model, as detailed in Eqs. (1), (3) and (6).

VI. CONCLUSIONS

ACKNOWLEDGMENT

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.
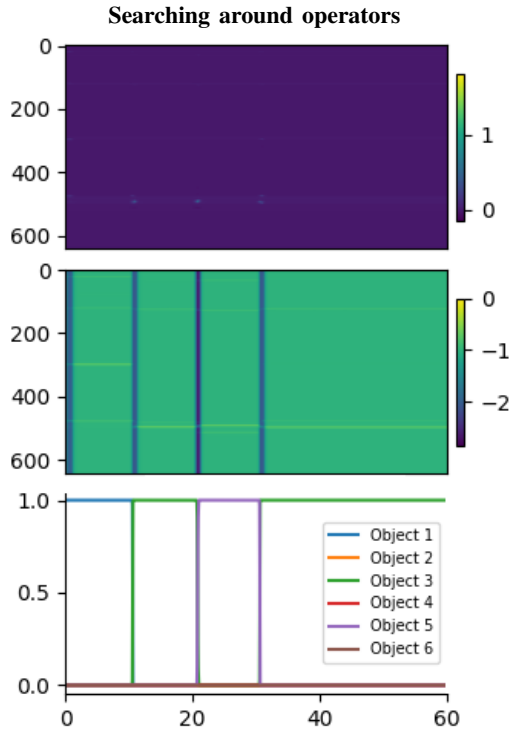
Fig. 5. In the experiment, the operator where activated at times 10, 20, 30 seconds, in the order *left*, *left*, *right*, *left*.

[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.