# JA-EGO: The Egocentric Dimension of Joint Attention in HRI

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

***Abstract*— Despite important progress in recent years, there is a long way to go for including social robots in our environments, capable of adaptation, interaction and communication. Our research is concerned with the study of social cognition in HRI, in particular 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 investigate the notion of egocentric localization and propose a neural network representation suited for joint attention research named JA-EGO. Inspired by *dynamic fields theory*, the model consists in a dynamical system representation which fuses information from immediate sensation and provides the means for attention selection and working memory under the influence of top-down and bottom-up modulation processes of attention in HRI. We firstly studied the attention selection model in simulations application scenarios. We then conducted a real experiment with the robot Pepper considering egocentric sources of information and basic proxemics. Results showed that JA-EGO is a convenient representation for HRI situations 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 reasoning and intelligence (like playing checkers), 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. Recent research has contrasted from one side 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* [5].

From the perspective of 4E cognition research, 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 skills so constituting forms of social attention at distinct levels of interaction and knowledge sharing [8].

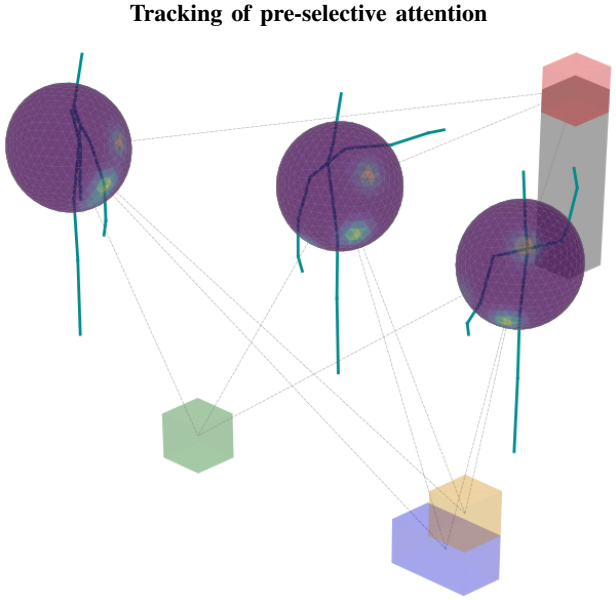


Fig. 1. The scene shows a situation where two agents are interacting about the orange object, whereas another agent is leaving the scene. Three other objects are present the effect of each object agents ego-spherical localization at a pre-selective stage is shown.

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* [4], 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).

From the considerations above, we propose the model named JA-EGO for tracking the attention focus of agents as represented from egocentric perspective, and resulting from acquisitions from robot's on-board sensory of agents body posture and local stimulus. For this, we model the evolution of attention selection as a dynamical system process represented by a neural network inspired on *dynamic neural fields* theory [1], which tracks egocentric focus of attention from spherical localization references.

[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.

This document is organized as follows: Section II discusses previous works and contextualizes our contribution according to limitations in state of the art research. Section III presents the mathematical definition of the model and discusses theoretical assumptions behind it. Section IV describes the methodology of the work, consisting in a preliminary study conducted in simulation for evaluating attention tracking based on the selection process, followed by an experiment with the robot Pepper considering a joint attention interaction situation based on egocentric sources of information and basic proxemics about objects in the environment. Section V reports the study's results, and Section VI presents conclusions and future perspectives.

## II. PREVIOUS WORK

Some works in the field of robotics have achieved impressive results by exploring attention saliency in sensory egocentric representations (e.g. for bottom-up [7] and top-down [2] mechanisms). Most of these works have considered a sort of environment exploration task, so the robot can focus on learning new things based on novelty. Fewer research have studied joint attention tasks (e.g. [2]) inspired on psychological theories of attention. Overall, the representation proposed has neglected bio-inspiration on neural systems, consisting mostly in storage arrays for data indexed by spherical tessellation mapping (see [6]). As a result, the dynamics of pre-attentive processing has not being modeled as a process unfolding in the same space where attention selection is done. We believe that this is an important limitation, when considering the possibility of investigating compositionality in joint attention as a descending (top-down) generative process combined with an ascending (bottom-up) saliency process, susceptible of study as a dynamical system.

Another limitation of previous research is considering the robot as the only one in interaction given with embodied ego-sensory mapping representations, so data coming from the human is mostly represented in the robot ego-sensory sphere. In our opinion, this would be a too much egocentric view of joint action in HRI. Thus, the robot should be able to represent its own world while accepting the egocentric view of others and being able to handle such body correspondences dynamically.

From our perspective, different works have constituted previous steps in the direction of proposing our current study which is worth mentioning. The work by [3] has proposed a sensory ego-cylindrical information fusion mechanism for egocentric localization allowing the robot to autonomous position with respect to object in the environment based on embodied representations. Concerning joint attention modeling, the model TOP-JAM [4] was proposed for interaction situations mediated by objects where allocentric references for addressing the task would make sense (e.g. sharing attention around a table, participating in an assembly collaboration task, to name a few) considering an extended range of knowledge and attention sharing (i.e. individual, monitoring, common, and sharing [8]).

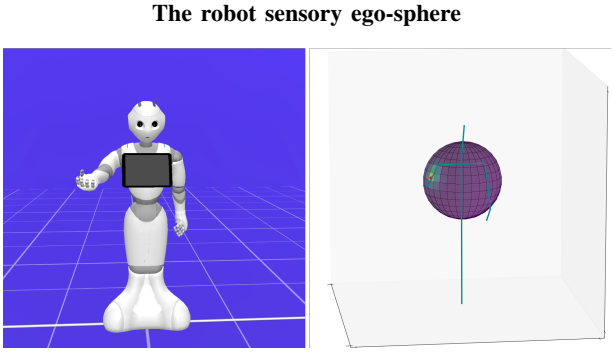


Fig. 2. Left: the robot is pointing to a location in the space. Right: the attention state of the robot is shown as the activation of the neural network representing the sensory ego-sphere, as stimulated by the intersection of the forearm direction with by the sensory ego-sphere.

## III. THE MATHEMATICAL MODEL

Let a ego-sphere representation of sensory information be modeled by the following neural network architecture [1].

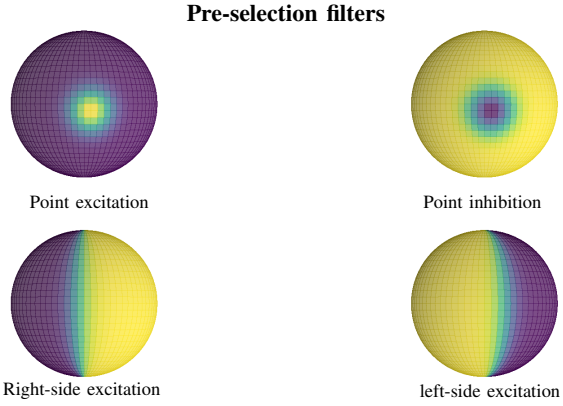


Fig. 3. Filters functions affecting the pre-selection phase.

## IV. METHODOLOGY

## 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).

## REFERENCES

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

[2] 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.

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

[4] 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).

[5] 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.

[6] 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.

[7] 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.

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