Attentional Mechanisms for Socially Interactive Robots – A Survey

João Filipe Ferreira, *Member, IEEE,* and Jorge Dias, *Senior Member, IEEE*

***Abstract*—This review intends to provide an overview of the state of the art in the modelling and implementation of automatic attentional mechanisms for socially interactive robots. Humans assess and exhibit intentionality by resorting to multisensory processes that are deeply rooted within low-level automatic attention-related mechanisms of the brain. For robots to engage with humans properly, they should also be equipped with similar capabilities. Joint attention, the precursor of many fundamental types of social interactions, has been an important focus of research in the past decade and a half, therefore providing the perfect backdrop for assessing the current status of state-of-the-art automatic attentional-based solutions. Consequently, we propose to review the influence of these mechanisms in the context of social interaction in cutting-edge research work on joint attention. This will be achieved by summarising the contributions already made in these matters in robotic cognitive systems research, by identifying the main scientific issues to be addressed by these contributions and analysing how successful they have been in this respect, and by consequently drawing conclusions that may suggest a roadmap for future successful research efforts.**

***Index Terms*—automatic attentional mechanisms, multisensory active perception, socially interactive robots, joint attention, bottom-up influences, top-down influences, probabilistic, hierarchical/modular architecture**

I. Introduction

IN the 2012 film "Robot & Frank" [1, 2], the main character, a retired cat burglar called Frank, is experiencing increasingly serious mental deterioration and dementia due to old age. Understandably, Frank's close family and friends fear he can no longer live by himself. Frank's son Hunter, after considering placing him in a nursing home, ends up choosing a different option: against the old man's wishes, he buys Frank a walking, talking humanoid robot programmed to improve his physical and mental health through a fixed daily routine that includes cognitive enhancing activities, such as gardening. Initially uncomfortable with the robot's presence in his life, Frank eventually warms up to his new companion's social skills, albeit not for the most honourable reasons – representing perhaps the most fanciful side of the film, the robot is eventually coached by Frank to help him return to his previous criminal ways.

J. F. Ferreira is with the Institute of Systems and Robotics and the Faculty of Science and Technology, University of Coimbra, 3030-290 Coimbra, Portugal (e-mail: jfilipe@isr.uc.pt).

J. Dias is with the Institute of Systems and Robotics and the Faculty of Science and Technology, University of Coimbra, 3030-290 Coimbra, Portugal, and also with the Khalifa University of Science, Technology and Research, Abu Dhabi 127788, UAE.

Manuscript received ..., 2013; revised ..., 2013.

Although there have been many attempts since the 1980s to produce convincing solutions in service robotics, this field has become increasingly relevant in the past decade. During this time, major technological hurdles have been conquered, and robots have become more and more affordable. Consequently, service robots have become a major part of assisted living technology [3, 4], even in its broader sense of giving assistance to promote the welfare of any human being, and not just elderly or people with physical disabilities. On the other hand, service robots have also been finding space in entertainment [5], and in assisting or replacing humans in tedious or difficult tasks [6], namely in industrial applications [7]. The fuzzy distinction between this type of robots and all others implies that service robots can range from devices with very limited cognitive skills, such as iRobot's Roomba™ [8], to the so-called "socially interactive robots" [9]. In other words, robots with varying degrees of social skills [4, 9], such as the robot portrayed in "Robot & Frank".

Most of the synopses used to promote this film include a small and apparently inconsequential remark on how its plot is "set in the near future". However, this subtle suggestion brings to the forefront two important facts:

1) The technology that will drive this kind of scenario in the "near future" is already being tested in setups simulating real-world applications (and hence the appropriateness of the word "near" to signify "close to the present day"). Numerous research projects, up until very recently or even at the time this text is being written, are deploying robots with social capabilities to interact with the final end-users in scenarios similar or somehow related to the situation found in "Robot & Frank", in order to test the skills and the overall acceptance of these types of robots [10] – for examples of projects sponsored by the EC (European Committee), see [11–17]. The fact that the film was ever made reflects how the general public is being made aware of the existence of such robots: the film's director, Jack Schreier, and writer, Christopher Ford, were reportedly inspired by present-day caretaker robots.
2) Human-robot interaction (HRI) technologies and corresponding cognitive capabilities of robotic systems have seen many developments in the last few decades – see for example [18] for an extensive survey. However, HRI technologies are still far from providing the degree of social capabilities to rival a human. On the other hand, strategies for dealing with proxemics (e.g. acceptability, etc.) and also safety issues regarding the robot autonomy

in a human environment are also still a matter of concern [3, 18]. These facts restrict most of the current socially interactive robots to highly controlled environments and specialised applications (and hence the appropriateness of the word "near" to signify "still not in the present day").

When interacting in socially-relevant applications, robots are expected to engage with humans while displaying attentional behaviours that resemble those of their interlocutors; as a matter of fact, they are supposed to be able to assess intentionality and to be, themselves, intentional agents. Humans assess and exhibit intentionality using processes that are deeply rooted within low-level automatic multisensory attention-related mechanisms of the brain – the popular saying "there is no second opportunity to make a good first impression" is not at all accidental. It is therefore the logical conclusion stemming from this reasoning that for robots to engage with humans properly, they should also be equipped with similar mechanisms.

Most of these mechanisms surface, in humans, during infancy, through interaction with their progenitors or caretakers. The acquisition of these basic skills plays a major role in the course of attaining advanced communication skills for more complex social interaction, in a process that takes place throughout the first twenty-four months of a child's life [19]. This fact has impressed the idea on many robot designers and researchers in artificial intelligence (AI) that robots should follow a similar route, engaging in sessions of what is called *learning by demonstration* to acquire equivalent mechanisms, from a "blank slate" starting point. This is part of the *epigenetic* [20, 21] and *developmental* [22] perspectives to robot design and AI, which is easily identifiable in many of the publications that we will be reviewing in this text. However, we argue that, although the origin of the robot's basic social skills is certainly relevant, even more important is the potential of an *emergent property* of more complex skills from basic building blocks, irrespectively of if these skills are learnt or preprogrammed (as an example of position papers that agree with this point of view, see, for example, [23]). Consequently, this is the perspective that we will be taking in this overview.

This survey thereby intends to provide a deeper insight in the modelling and implementation of automatic multisensory attentional mechanisms for socially interactive robots. This will be achieved by summarising the contributions already made in these matters in robotic cognitive systems research, and by consequently drawing conclusions that may suggest future successful research endeavours by analysing whatever scientific issues still remain unsolved.

II. Contextualisation

We shall start by introducing the basic definitions, followed by a presentation of the concepts and paradigms justifying our survey.

A. Basic definitions

Attention is the process whereby an agent allocates perceptual resources to analyse a subset of the surrounding world in detriment of others [24–26]. This allocation might consist of an explicit reorientation of sensors (in the visuoauditory case, *head reorientation*) to a specific part of the environment, in which case we are talking about *overt attention*, or the deployment of computational resources for processing a specific part of the stream of sensory information, in which case we are referring to *covert attention*. Overt attention, therefore, is the direct cause of the process called *active perception* [27, 28]; it can be either voluntary or involuntarily driven, and in the latter case, active perception is characterised by the *automatic orientation* of sensors. When referring specifically to vision, this is implemented by the conjunction of *gaze shifts* based on head reorientations and sequences of *saccades* (i.e., rapid eye movements) and *fixations*.

It has been found that these two types of attention are mostly used in conjunction with one another (mainly, overt processes follow covert mechanisms in the overall timeline for attention), although both might operate independently [24]. The "centre of spotlight" of attention, be it covert or overt, is usually referred to as the *focus of attention* (FOA). Although intuition may lead us to believe that we are able to perceive our surroundings mostly in a single sensory snapshot, in fact only the FOA is analysed in detail, while the rest is a result of accumulating perceptual information in short-term memory. In fact, regions other than the FOA are mostly ignored at any specific instant, causing a phenomenon called *change blindness* for peripheral stimulation [24].

Despite the obvious disadvantages of side-effects such as change blindness, the fundamental reason for the existence of attention is easily understood: it otherwise exponentially increases the efficiency of perception. This increase is obtained in two ways:

- An immense amount of information is relayed by sensors to the perceptual resources of a cognitive system at any given instant – for example, in vision, this corresponds to a stream of data of around 10^8 to 10^9 bits per second [25]. Processing all this information in real-time is intractable, and hence attention serves to ensure that only the most *relevant* data (we will address the question of relevance below) is actually processed.
- On the other hand, it is important for a perceptual system to deploy resources in a inhomogeneous way, so that most of the processing power and resolution is directed to the centre of the focus of attention and therefore allowing for a lower resolution in the periphery (in the case of vision, this is called *foveation*, relating to the fovea, the centre of the retina, the area of greatest density of photoreceptors in the eye).

Additionally, attention is triggered by two different types of influences – it might be *stimulus-driven* (i.e., affected by *bottom-up influences*), or *goal-directed* (i.e., fuelled by *top-down influences*) [24–26].

So, given that apparently we are only talking about perception and action for perception, what has attention to do with social interaction? Overt attention and active perception, in particular, have a deep significance for social beings – they are the *precursor of intent*. Since attention is influenced by the agent's goals, its overt manifestations reflect them as an

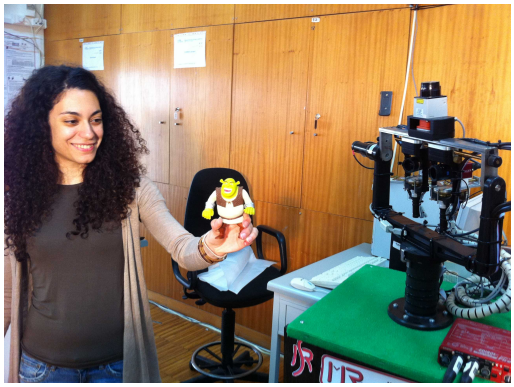

Figure 1. Joint attention in the context of social interaction with robots, using the Institute of Systems and Robotics' Integrated Multimodal Perception Experimental Platform (IMPEP). Hardware and motors for this platform were mounted within the scope of the Perception on Purpose (EC project number FP6-IST-2004-027268) project, and sensors were installed within the scope of the Bayesian Approach to Cognitive Systems project (EC project number FP6-IST-027140).

action and allow an external observer to get an inkling of these *intentions*. As Kaplan and Hafner [29, 30] state, overt attention is, in fact, *intentionally directed perception* – see also the concepts of *intentional contingency*, explained by Blakemore et al. [31], and *intentional stance*, referred to by Scassellati [32].

When two interlocutors look at each other (i.e., they become the focus of attention of one another), we have what is called a *dyadic agent-agent relation*. However, when one of the agents changes his FOA intentionally to a third entity, and the second agent follows by *acknowledging* this intention and redirecting his own FOA to that same third entity, we are now in the presence of a *triadic relation*. This intentional attention coordination together with mutual awareness is called *joint or shared attention*. In the human developmental timeline, this is the first fundamental skill for social interaction relating to attention [22, 29, 30, 33, 34]. Consider the following illustrative example:

> A mother looks at a toy, looks back at her two-year-old child to make sure that she looked at the toy, and utters the word "elephant". Then, the child nods and repeats the word "elephant", thus signalling the understanding of her mother's intention. A few days later, the child exchanges roles with the mother, while trying to make her understand that she wants her to play with another object, in this case a toy dog.

Joint attention is shown in the context of social interaction between a robot and a human interlocutor in Fig. 1. As an interactive process, it is dynamic, cyclical and dividable into several phases: *initiating joint attention* (IJA) by the agent through attention manipulation; *responding to joint attention* (RJA) through attention detection, by resorting to gaze following or through the detection of other cues such as pointing; and, as recognised by some authors, *acknowledging joint attention* (AJA) and the more proactive *ensuring joint attention* (EJA), which are the processes of monitoring and re-establishing the IJA-RJA cycle.

Some authors make a distinction between the concepts of joint and shared attention, by considering one as the mechanical manifestation of the other (i.e., coordination without considering awareness – for a discussion on the ongoing debates over these definitions, see for example [35]). Although this distinction is important – we will, in fact, address it in the context of this survey – for the sake of clarity and conformity with cited work, we will use both designations interchangeably.

As can be understood by the previous example, when considering joint attention, we are in fact talking about a *joint action*. Joint actions, as clarified by Fiebich and Gallagher [36], are coordinated behaviour patterns, specified depending on the shared intention of the agents, of which they are fully aware. This implies, not only a shared goal, but also a common goal in what Fiebich and Gallagher call a "we-activity" – an activity shared by both agents. It is therefore easily acknowledgeable that this is the foundation of social interaction, for either biological or artificial agents.

B. Background and Motivations

The human perceptual system includes fast processing lanes committed to automatically producing a quick description of the environment. This description is then used to discriminate between features of incoming stimuli in order to establish *relevance* and hence prioritise further processing – in other words, to decide the next focus of (overt and/or covert) attention. Complementing these lanes, many involuntary[1] mechanisms driven by attention elicit automatic motor reactions to relevant stimuli (e.g., automatic orientation). The significance of such lanes have led many authors to argue that preliminary descriptions might be fed back so as to improve the results of slower perceptual processes (see, for example, [37]), and also that the corresponding involuntary and automatic mechanisms have a significant impact on how humans behave and react to to the surrounding world (see, for example, [38]).

As shown by Corbetta and Shulman [26, see section on "Stimulus-driven control of attention", pages 207 and 208] and more recently by Corbetta, Patel, and Shulman [39], involuntary attentional mechanisms, while apparently fundamentally stimulus-driven, are also modulated by goal-directed influences through the so-called *attentional sets* that impose *task relevance* as a prioritising measure. The sensory (bottom-up) distinctiveness of a feature interacts with the ongoing cognitive (top-down) goal. In fact, distinctive sensory stimuli (the measure for "distinctiveness" was dubbed by Koch et al. [40, 41] in their seminal work as *saliency*) attract attention more effectively when they are relevant to the task at hand, in other words, when they are *contingent*. On the other hand, these authors show that there is another system involved in involuntary responses to salient or unexpected stimuli, described as a system that is specialised in the detection of *behaviourally relevant* stimuli. It is therefore intuitive to conclude that, for social beings such as humans, *socially relevant stimuli are generally both behaviour- and task-relevant*. This becomes

[1] But not necessarily unconscious.

particularly difficult to refute when studying attention in the course of social interaction.

Fiebich and Gallagher [36] build up a strong case for studying joint attention as a fundamental part of social interaction. In fact, they propose an instrumental account of basic joint action and proceed to argue that *intentional joint attention is a basic joint action*, and as such it is involved in most complex joint actions, being substituted in very exceptional cases (mostly when face-to-face interaction is not possible) by a certain use of language. Actually, joint attention is the precursor of many fundamental types of social interactions, such as *imitation* or even *social referencing*, the process whereby the agent seeks out emotional information from the interlocutor, "in order to make sense of an event that is otherwise ambiguous or beyond that individual's own intrinsic appraisal capabilities" [42]. Additionally, Scassellati [32] argues that if a robot is supposed to be able to engage in social dynamics, it must possess, as humans, a set of metarepresentational abilities that would allow it to understand the intentional stance of its interlocutors. These abilities have been collectively called a *"theory of mind"*, and are reviewed in the context of robotic cognitive system design by Scassellati [32] in his seminal publication. This concept, originally stemming from philosophy and psychology, is supported by findings in neuroscience regarding the existence of *"mirror neurons"* that are instrumental for observers to infer the intention of others. Scassellati then proceeds by showing that some of these abilities subsume gaze following and joint attention. Although none of the two "theory of mind" models analysed by the author is shown to be exactly suited for robotic implementation, Scassellati demonstrates that such an implementation would be an important contribution to socially-capable robotics. In conclusion, joint attention presents the perfect opportunity to experience attentional mechanisms in action in the context of social interaction.

On the other hand, as referred in the previous section, *attention is a multisensory experience*[2]; as such, although vision has been by far the most studied sensory modality in studying attention (and, by consequence, joint attention), it should always be considered in the context of the interaction between more than one sensor, most notably vision and audition, the two quintessential noncontact sensors in humans. Other sensors indirectly involved in socially-relevant attention processes would be, for example, those relating to proprioception (i.e., sensors that provide feedback on one's pose).

Finally, all attentional mechanisms inherit the *uncertainty* congenital to both perceptual and actuation processes; therefore appropriate modelling and computational tools should be applied in their development in robotic cognitive systems.

Therefore, we find that an ideal approach to the study of the influence of *automatic multisensory attentional mechanisms in the context of social interaction with robots* should be regulated by the following agenda:

- *How do these mechanisms influence joint attention?*
- *What is the interplay between the automatic mechanisms underlying multiple sensory modalities and actions in the course of joint attention?*
- *How does one deal with the uncertainty inherent to attention?*

Consequently, our article may be construed as a pertinent follow-up to the just over half-a-decade old seminal sequence of publications by Kaplan and Hafner [29, 30], although aiming at different surveying objectives. As will be further detailed in section IV, this review focusses only on the scientific issues raised by the *automatic mechanisms* involved in joint attention. Also note that, as we mentioned in the introductory section, we are not as interested in the developmental temporal sequence of emergence of complex social skills as we are in the *hierarchical emergence of complex social behaviours* that underpin adaptive social interaction (i.e., a complex joint action is more than just the sum of its basic joint actions).

Therefore, in the following sections we will proceed by introducing the main models and approaches to joint attention implementation in robots, followed by a discussion of the main scientific issues that relate to our motivations and the above-mentioned agenda, and wrap up by drawing conclusions regarding these matters.

III. Models and Approaches

A. Brooks, Scassellati and Breazeal

The work surveyed in this subsection is the result of the effort of a group of researchers led by Rodney Brooks who have driven one of the three main pioneering endeavours in modelling joint perception, and closely related processes, such as imitation.

This endeavour received its first major impulse with Brian Scassellati's manifest for the development of humanoid robots including (at least parts of) a "theory of mind" [32]. Following the ideas represented in this manifest, several frameworks were developed around well-known robotic platforms, starting with the Cog upper-torso humanoid robot [46–48]. For this particular robot, Brooks, Scassellati et al. mainly address issues related to attention detection.

Next, Breazeal and Scassellati [49–52] teamed up to develop the Kismet active head robot (Fig. 2), taking Scassellati's paradigm one step further by adding a model for the emergence of social coordination, integrating perception, attention, drives, emotions, behaviour arbitration, and the emulation of expressive acts. As the corollary of their joint work, the two researchers wrote one of the most seminal publications on robotic imitation [53].

Cynthia Breazeal then went on to continue studying social interaction with robots through another platform – a doll-like robot with marginally humanoid features called Leonardo (or Leo). With this platform she added more emphasis to the study of the emergence of intentional understanding [54–56].

Scassellati, in the meanwhile, continued to work separately on joint attention and related mechanisms – see, for example, [57], where Doniec et al. use a robot named Nico to study the influence of imperative pointing and reaching in the learning of joint attention.

[2]Actually, attention processes have been found to exist for virtually all types of human sensing – for auditory attention, see for example [43]. Tactile sensing is another representative example [44], having already spurred similar solutions in robotics [45].

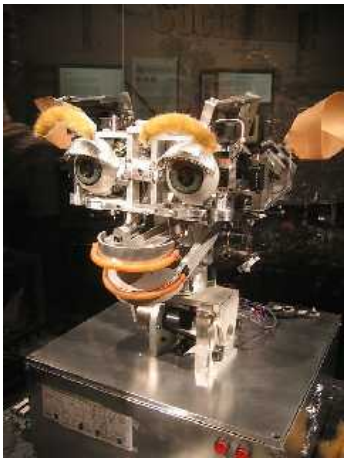

Figure 2. Kismet, the late 1990s active head robot built at the Massachusetts Institute of Technology (MIT) by Cynthia Breazeal and colleagues, now residing at the MIT Museum (© Jared C. Benedict / Wikimedia Commons / CC-BY-SA-2.5).

B. Asada, Ishiguro, Nagai and Sumioka

Another pioneering group was embodied by the team originally teamed up by Minoru Asada, roughly beginning with the group's cognitive developmental robotics (CDR) paradigm proposal [58], which, according to the authors, "aims to understand the cognitive developmental processes that an intelligent robot would require and how to realize them in a physical entity".

Within this group, Hiroshi Ishiguro tried to address multisensory attention and multimodal human-robot communication through work supported by the Robovie robot [59, 60]. Imai, Ono, and Ishiguro [61] took advantage of this platform to devise a framework able to attract a human's attention by pointing at an object and establishing mutual gaze.

Yukie Nagai turned the group's focus towards the developmental process of attention detection, instead of just devising a framework simply capable of performing it (Fig. 3). Nagai et al. [62–65] present a framework that learns the correlation between the gaze of a human and an object in the visual field at a certain position, namely by investigating the role of motion information in this process (resorting to the concept of *motionese*, a stimulus-driven attentional feature) [66, 67]. The robot progressively learns to estimate human gaze direction in order to find objects more rapidly, thereby acquiring a gaze following skill.

Motion and also *transfer entropy* were also investigated by Asada's research team through Sumioka et al. [68–71], although only in [68] an actual physical robot was used, while the rest were simulated experiments. Transfer entropy is an information theoretic measure used to detect causality that shares some of the desired properties of mutual information but additionally takes into account the dynamics of information transport.

The most recent research efforts by this group include studying joint attention from a multisensory perspective [72] and joint attention coordination through contingency [73] (again, in simulation only), including IJA, RJA and AJA.

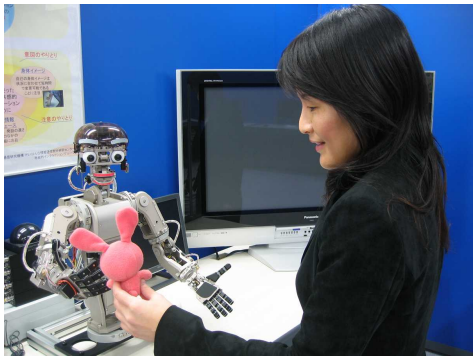

Figure 3. Yukie Nagai shown demonstrating the performance of her group's developmental learning process for joint attention during interaction with a robot [63, 65] – reproduced by kind permission.

Figure 4. Frederic Kaplan shown holding an AIBO robot, a model used in a set of pioneering studies in robotic joint attention by his research team [74] – reproduced by kind permission.

C. Hafner, Kaplan and Oudeyer

Another important research team, fronted by Frédéric Kaplan and Verena V. Hafner, although arguably not as prolific in the development of robotic applications specifically relating to joint attention as the previous groups, has nevertheless arguably produced the most influential sequence of surveys on this subject to date [29, 30], helping to frame most of the research work in this area from 2004 onwards (Fig. 4).

As a first example of research work by this group more closely related to the study of joint attention processes in robots, Hafner and Kaplan [74] showed how four-legged robots can learn to interpret each other's pointing gestures. In this work, one of the robots took the role of a caregiver/teacher pointing to an object, and the other, the learner, would then have to interpret the pointing gesture correctly in order to find the object. At more or less the same time, Hafner and Kaplan [75] and Kaplan and Oudeyer [76] investigated issues related with the emergence of the self and with the identification with others. The objective was to find internal abstract measures allowing for a distinction between autonomous behaviour and coupled interactions with peers.

D. Other Approaches

Ever since the pioneering efforts by the aforementioned research groups, either in parallel with these endeavours or inspired by their success, a great deal of research has followed

the interest for the study of joint attention and related mechanisms. In the following text, we will present a transversal review of what we feel to be a representative sample of this body of work, with the primary aim of surveying research in robotics and computational modelling that will allow us to follow the agenda defined in section II-B.

In 2001, Kozima and Yano [77, 78] designed a robot called Infanoid (also used later on by Nagai [66]) with the ability to track faces and objects with salient colour, to point and reach for those objects, and to alternate gaze between faces and objects.

Haasch, Hofemann, Fritsch, and Sagerer [79] presented a multimodal object attention system capable of identifying objects referenced by the user with gestures and verbal instructions, and tested it using the BIRON robot companion [80] devised in the scope of the COGNIRON European project [14]. Their system was able to detect known and unknown objects and would store newly acquired object information in a scene model for later retrieval. This way, the growing knowledge base of the robot companion would improve the interaction quality as the robot could more easily focus its attention on objects it had been taught previously.

Hoffman, Grimes, Shon, and Rao [81, 82] presented an elegant solution, closely related to the approach proposed by Breazeal et al. [51] for Kismet and Leonardo in the sense that they used stimulus-driven and goal-directed attention mechanisms to fuel a gaze imitation process for joint attention. However, Hoffman et al.'s approach differs from the latter in its emphasis on the use of a unifying probabilistic framework at all levels, producing a maximum a posteriori (MAP) estimate of the object attended to by the human instructor from three different vision-based sources: (1) a model-based algorithm for estimating an instructor's gaze direction, (2) a bottom-up image saliency algorithm for highlighting behaviourally-relevant regions in the image, and (3) a top-down saliency map that biases the imitator to specific object preferences of the instructor, as learned over time.

Yücel, Salah, Meriçli, and Meriçli and colleagues [83, 84] presented a cognitively-inspired, virtually task-independent gaze-following/fixation and object segmentation mechanism for robotic joint attention. In their first approach [83], the head pose of the caregiver was determined by the proposed system and gaze direction estimated from it. At the same time, the depth of the object along the direction of gaze would be inferred from head orientation. The intersection of gaze direction and depth provided the system a coarse estimate for centre-of-mass of the object. Finally, using salient features and pooling a number of estimates, the final segmentation of the fixated object would be performed. In their most recent work [84] (Fig. 5), Gaussian process regression and neural networks were contrasted to interpolate gaze direction. Then, the authors combined gaze interpolation with image-based saliency to improve the target point estimates, and in the process tested three different saliency schemes, the three of which were shown to improve object location estimation when compared to using regression alone. They further concluded that it was not possible to "single out any of the saliency schemes as being clearly superior to the others". This work was experimentally

Figure 5. Experimental setting of the work by Yücel et al. [84] – reproduced by kind permission. To evaluate their method, several experiments were performed to model joint attention between a human experimenter (in the photo, on the left, co-author Cetin Meriçli is shown acting out the human experimenter's role) and an embodied agent. In this setting, the experimenter and the robot place themselves roughly 1.5 meters apart at opposite sides of a table, on which a number of objects are placed in a non-occluding fashion. A session of joint attention is initialised when the experimenter establishes eye-contact with the robot, which leads to a fully-frontal face image being acquired. The experimenter fixates his/her attention to one of the objects by looking at them in random order for a certain duration of time.

supported by an Aldebaran Nao humanoid robot [85] as the main interaction unit and the FESTO Robotino robot [86] for navigation.

As opposed to simply designing a system with a preprogrammed modular ability of attention detection, Carlson and Triesch [87] conceived a computational model to deal with the emergence of gaze following through reinforcement learning. In the course of their research, the authors identify a basic set of mechanisms which they claim are sufficient for the development of this skill.

Andry, Gaussier, Moga, Banquet, and Nadel [88] proposed a neural network (NN) architecture designed to exhibit learning and communication capabilities via what the authors claim to be a "proto-imitation" behaviour dealing with (and, according to the authors, taking advantage of) the ambiguity inherent to perception. This approach is implemented in the perspective of studying turn-taking and gestural communication between two agents. Synchronisation is obtained as a global attractor depending on the coupling between agents' dynamics. Additionally, the authors discuss the non-supervised context of the imitation process by presenting experiments in which the same architecture is able to learn perception-action associations without any explicit reinforcement. Learning is supported by the ability to detect novelty or irregularities in the rhythm of communication. Ten years later, Boucenna, Gaussier, and Hafemeister [89] followed up by designing a system modelling joint attention for social referencing by using gaze following, facial expression recognition and motion detection mechanisms (the first two learnt autonomously) to associate an emotional value to an attended object. The authors claim that this framework is complementary to the developmental model proposed by Nagai et al. [63].

Following a similar line of interest, Ikegami and Iizuka [90] proposed a coupled dynamical recogniser as a model for simulating turn-taking behaviour. A recurrent neural network

(RNN) is used to produce the motor outputs of mobile robot agents with two wheels, which then compete in simulation to take turns on a two-dimensional arena. By using a genetic algorithm technique, the authors show that turn-taking behaviour is developed between two agents, which is established even in the presence of a variety of dynamics.

Ito and Tani [91–93] presented research on the engagement of a humanoid robot with humans (more specifically, the Sony QRIO SDR-4XII [94]), including stable and unstable phases of interaction using particular entrainment imitation dynamics. More specifically, they proposed an extended scheme of RNNs for constructing a mirror system by which recognition of the other's movements and generation of self-motion can be naturally synchronised in the real-time imitation processes in the context of joint attention.

The research team of Roseli Romero [95, 96] proposed an architecture that simulates an individual's operant conditioning (a type of learning in which an agent's behaviour is modified by its consequences) through histories (sequences) of reinforcement. It is composed by three main modules: (1) a stimulus perception module, working upon a vision system based on the work by Breazeal and Scassellati [49] described earlier; (2) a response emission module, composed by a learning mechanism that constructs a nondeterministic policy for response emission, in other words, to determine what response is to be emitted in the presence of a certain antecedent stimulus; and (3) a consequence control module, composed by a motivational system that simulates the intrinsic needs of the robot and detects reinforcements received from the environment. The motivational system is formed by "necessity units" that are implemented as a simple perceptron with recurrent connections (similar to a RNN), thereby, simulating the homoeostases (self-regulating processes to achieve metabolic equilibrium) of living organisms. More recently, the same group of researchers compared the contingency learning process that they originally proposed with two other approaches [97], ultimately determining experimentally that, between the three, an extension of the TG relational reinforcement learning algorithm [98] yielded the best performances. Experiments throughout this body of work were conducted using the robotic head—WHA8030 of Dr. Robot.

While most of the approaches mentioned so far have focussed on dealing with responding to joint attention through attention detection, and considerably less on the process of initiating joint attention by the robot through attention manipulation, very few exceptions study the relevance and implementation of mechanisms for ensuring joint attention. A rare example would be the work by Huang and Thomaz [99, 100], where the authors propose a framework that models IJA, RJA and EJA, tested using the upper-torso robot Simon. In the proposed model, RJA and IJA run exclusively. Conversely, EJA is a permanently ongoing monitoring process that coordinates with IJA to ensure that the other agent attends to the right focus. The authors perform an experimental evaluation that allows them to ultimately conclude that RJA and EJA both contribute to a greater acceptability of the robot by the human interlocutor, and moreover argue that EJA improves the robot's performance in interactive tasks.

IV. Discussion of Targeted Scientific Issues

Kaplan and Hafner [29, 30] identified four prerequisites for joint attention: (1) the ability to track the observable attentional behaviour of other agents (*attention detection*); (2) the capability of influencing the FOA of other agents (*attention manipulation*); (3) the ability to engage and regulate social interaction (*social coordination*); (4) the ability to view oneself and the other as intentional agents, and as such to acknowledge each other's intentions (*intentional understanding*). These prerequisites are *intrinsically and fundamentally related to automatic attention-related mechanisms*. In reality, this happens to a point where not only perception of the other's intent is modulated by these processes, but also self-goals are intentionally relayed to the other by taking advantage of the subconscious knowledge of involuntary attentional reactions of the interlocutor. This fact is made particularly clear in the sociological study by Muhl, Nagai, and Sagerer [101], investigating how humans react to artificial agents in HRI and how this reaction is influenced by attention, and also by the work of Schillaci, Bodiroža, and Hafner [102], who study the effects of different attentional models during this kind of interaction.

Consequently, in this section we will analyse a total of twenty (20) of the models presented above[3] in the light of the agenda we established in section II-B.

We will start by assessing the use of state-of-the-art implementations of stimulus-driven preattentive and automatic overt attention processes in the context of robotic social interaction. We will then successively link these processes to each of Kaplan and Hafner's prerequisites:

- first, we will analyse how the authors of those models tackled the mapping between the robot's self, the other and the third entity involved in joint attention through prediction, thus providing a bridge between prerequisites (1), (2) and (4);
- second, we will evaluate how their approach to addressing prerequisites (3) and (4) affects goal, action and behaviour selection, and therefore modulates attention (i.e., through top-down influences).

We will end our discussion by reviewing experimental validation methods and criteria which indirectly but unavoidably shape how the scientific issues are targeted, since researchers will naturally try to achieve the highest scores possible for the criteria they select.

A. Preattentive and Automatic Overt Attention Processes for Social Interaction

The attentional mechanism that drives involuntary processes of overt attention and active perception (i.e., the *attentional capture*, resulting from the so-called *pop-out stimulus effect*) has been identified as the fast, automatic, *preattentive* evaluation of the relevance of the incoming stimuli according to *basic features*, which is then encoded into sensory-centred spatially

[3]More concretely, analysis is performed by assessing model relevance case-by-case, and only if an actual physical robot is used, or, if simulated, close to final implementation.

organised maps (such as *saliency maps* [41], which in the case of vision are said to be *retinotopically arranged*[4]) modulated by goal-directed influences [24–26, 103]. As discussed earlier, this evaluation is performed according to behavioural relevance; therefore, if the task is social interaction in the context of joint attention, behaviourally-relevant features should naturally include socially-relevant features.

Another fundamental issue is the *unit of attention* [24]: should one consider location, features or objects, or a conjunction of all of the above, as seems to be the current consensus? The use of saliency maps, in a way, provides the means for unifying the first two units, while the notion of grouping features to form volatile perceptual units called *proto-objects* – see [104, 105] – supports full conjunction of all the above. Attention, as a perceptual mechanism, is a *multisensory process*; therefore, it is also important, in the context of social interaction, to take into account visual, auditory and visuoauditory units of attention.

There is some debate on what constitutes the actual set of basic features (also called *primitive features or attributes*) [24]. Colour, luminance (either independently or associated to colour), orientation, size (including length and spatial frequency) and motion seem to be indisputably basic, while luminance onset (flicker), stereoscopic depth and tilt, shading, novelty, or, in audition, hearing one's own name, and many others are often proposed as primitive features, although not consensually [24]. Two particularly interesting saliency metrics – which are not at first glance biologically plausible nor are they actual basic features (in fact, they result from mechanisms that try to profit from the underlying statistics of basic features) – have nevertheless had great success in predicting human overt visual attention behaviours: *Bayesian surprise* (strongly related to novelty), introduced by Itti and Baldi [106, 107], and *information maximisation*, introduced by Bruce and Tsotsos [108, 109]. Additionally, a particularly important issue is if *faces*, the main vehicles of social intention display, are, themselves, basic features or not – either only detected as such (1), or individually recognised as belonging to different individuals (2), or even classified as conveying a specific emotion through expressions (3). Palermo and Rhodes [110] show that, apparently, (1) is basic, while (2) and (3) are probably not.

The relevance of each of these features is then weighted in terms of priority, as explained in section II-A, through attentional sets selected according to top-down influences. However, Corbetta and Shulman [26] demonstrate that in the human brain there are mechanisms that override the current attentional set when the agent is faced with unexpected and overbearing events (e.g. a loud noise, or a sudden flash of light). Additionally, recently attended regions of space are less likely to be re-attended in the near future – this is called *inhibition of return* (IOR) [24, 25], and it is debatable if it is actually a self-contained process or the manifestation of an underlying process or set of processes.

So, in a nutshell, stimulus-driven pop-out effects play a crucial role in joint attention:

1) they automatically draw perception towards features that potentially promote the initiation of joint attention;
2) they attribute task- and socially-relevant value to perceived stimuli, therefore expediting joint attention as an ongoing process;
3) they embody common ground for sharing self-intention and the intention of the other.

In the work of Kirchner and Alempijevic [111], the reader may find a very interesting account of the practical importance of attention detection (included in what the authors refer to as the robotic "Read" action/ability) and manipulation (similarly subsumed within the robotic "Elicit" action/ability) in a robot-centric perspective regarding HRI.

On the other hand, top-down influences, despite their obvious importance, are mostly neglected in existing models when compared to bottom-up influences, given the difficulty in concretely defining such cues [24, 25]. Among these influences, the hardest to conceptualise are precisely those with social relevance, such as the agent's emotions, desires and motivations [24] – we will address these specifically in section IV-C, given their importance for joint attention. Prior knowledge relating to *contextual information*, however, has recently been the subject of important research efforts – previous experience for a specific perceptual scene, or also the semantic category of that scene (the so-called *gist* [112]) are important examples of this type of cognitive, top-down influences [24].

Visual attentional mechanisms and broad computational frameworks were thoroughly discussed by Frintrop, Rome, and Christensen [24] (among which two particularly seminal contenders would be the work by Koch and Ullman [40] and Tsotsos et al. [113], which differ mainly on how they approach the computation of the final saliency map), while computational models of visual attention were extensively reviewed by Borji and Itti [25]; we highly recommend reading these groundbreaking publications for more information in the overall outlook on these issues. There is no systematic survey of mechanisms and computational models of attention for any other sensory modality, such as audition, as far as we are aware of.

The models presented in section III are compared in Table I in terms of preattentive and automatic attention capabilities, including the respective computational approach and the basic features used for each model, and also the supporting hardware and sensors. As can be seen in this table, preattentive and automatic mechanisms of attention have been extensively used to drive joint attention. However, not all mechanisms share the same popularity: in vision, for example, colour, motion-related and face basic features are used by 80%, 55% and 80% of the surveyed work, respectively, as opposed to 45% for luminance and orientation; auditory basic features are clearly less popular than their visual counterparts (obviously due to most solutions – 65% – being visual-only); finally, IOR is only modelled explicitly in very few occasions (10%).

[4]The reason for this is that the computational framework of the visual saliency map models what is believed to be a direct flow from the retina to an early representation, where features are extracted and represented in parallel [40], thus translating to computer vision implementations through a pixel-wise association of saliency values.

Table I
COMPARISON OF SURVEYED RESEARCH WORK IN TERMS OF HARDWARE, SENSORY CAPABILITIES, AND PREATTENTIVE AND AUTOMATIC ATTENTION MECHANISMS.

				Basic Visual Features[d]					Basic Auditory Features[e]				
Study	Platform[a]	Sensors[b]	Approach[c]	C	L	O	M	F	AL	HV	ON	MS	IOR
Scassellati [46, 48]	UT	VA+P	D	+	+	-	+	+	+	-	-	-	+
Breazeal and Scassellati [49–52]	AH	VA	D	+	-	-	+	+	-	+	-	-	-
Breazeal [54–56]	TR/HR	VA+P	D	+	-	-	+	+	-	+	+	+	-
Doniec and Scassellati [57]	UT	V	NN	+	-	-	-	+	-	-	-	-	-
Ishiguro et al. [59–61]	HR	VA	D	?	?	?	?	?	-	+	-	-	-
Nagai et al. [62–65]	UT	V	NN	+	+	-	+	+	-	-	-	-	-
Nagai et al. [66]	HR	V	NN	+	-	+	+	+	-	-	-	-	-
Nagai et al. [67]	HR	V	D	+	+	+	+	+	-	-	-	-	-
Sumioka et al. [68]	HR	V	NN	+	+	+	+	+	-	-	-	-	+
Hafner and Kaplan [74]	TR	V	D	-	+	+	-	-	-	-	-	-	-
Kozima and Yano [77, 78]	HR	V	D	+	-	-	-	+	-	-	-	-	-
Haasch et al. [79]	MR	VA	D	+	?	?	-	+	+	+	+	+	-
Hoffman et al. [81, 82]	AH	V+P	P	+	+	+	-	+	-	-	-	-	-
Yücel et al. [83]	HR	V	D	+	+	+	+	+	-	-	-	-	?
Yücel et al. [84]	HR	V	P *vs* NN	+	+	+	-	+	-	-	-	-	?
Andry et al. [88]	TR	V	D	-	-	-	+	-	-	-	-	-	-
Boucenna et al. [89]	AH	V	NN	+	-	+	+	+	-	-	-	-	-
Ito and Tani [91–93]	HR	V	NN	+	-	-	+	-	-	-	-	-	-
Romero et al. [95–97]	AH	VA	NN	+	+	+	-	+	-	+	?	-	-
Huang and Thomaz [99, 100]	UT	VA	D	-	-	-	-	+	-	+	?	-	-

In general: + implemented; - not implemented; ? not clear from author's description.

[a] UT – upper-torso; AH – active head; TR – toy robot; HR – humanoid robot; MR – non-humanoid mobile robot.
[b] V – vision-only; VA – visual and auditory sensors; P – proprioceptive sensors (only if explicitly used in the context of joint attention experiments).
[c] D – deterministic; NN – neural network; P – probabilistic; *vs* – comparative study (versus).
[d] C – colour; L – luminance; O – orientation; M – motion; F – faces.
[e] AL – auditory localisation of sound source; HV – human voice detection; ON – own name detection; MS – multiple sound sources.

Similarly, from Table I it becomes clear that deterministic (and mostly heuristic) solutions are the most prevalent computational approaches to modelling (55%), followed by neural networks (40%), and finally, from a significant distance, by probabilistic solutions (10% of the sample). We argue that deterministic solutions are of very little application value in unconstrained, realistic conditions, while neural network-based solutions are only marginally better; we posit, on the contrary, that probabilistic approaches would be the appropriate approach to deal with uncertainty, according to the desiderata we defined in section II-B.

The reasons for the current *status quo* in this respect are clear – solutions that attempt to be generic have acknowledged the importance of these mechanisms and approaches, while tailor-made solutions, constituting the noticeable majority, have used these tools only while they are useful for their purposes. For example, it is not a coincidence that colour is a highly popular basic visual feature. Most experiments in the literature involve identifying brightly coloured and contrasting objects, which evidently greatly simplifies segregation, but at the significant expense of the usability of the proposed systems in unconstrained, real-world, social HRI applications. In fact, our assessment of this survey is that, as time goes by, solution proposals resulting from research efforts in this area have consistently been more and more specific. We conjecture that the complexity of modelling joint attention has led researchers further away from acknowledging the potential of automatic attentional mechanisms and probabilistic approaches to modelling.

B. Automatic Anticipatory and Predictive Mechanisms for Closing the Action-Perception Loop – Going from Dyadic to Triadic Representations

In neuroscience, the *ideomotor principle* (IMP) [114] emphasises the importance of anticipating (implicitly or explicitly) the sensory consequences of our actions and the actions of others for: 1) adaptive behaviour, 2) guidance of attention, 3) mentalising abilities and 4) social learning, all of which are a powerful means for building artificial cognitive systems that can acquire new knowledge autonomously, that learn from humans or that adapt to particular environments and preferences of the users [53, 115].

In the context of joint attention, three requisites need to be fulfilled so as to apply the IMP:

1) the ability to predict the sensory consequences of one's own actions through *egocentric representations*;
2) the ability to predict the sensory consequences of the other's actions through *dyadic representations*;
3) the ability to establish triadic correspondences by relating egocentric and dyadic representations of the third

entity.

A substantial amount of robotic cognitive systems for joint attention typically use sensory-centred saliency maps, already aiming to provide an egocentric representation. Traditional two-dimensional saliency-maps, however, are related to vision, and as such have been modelled retinotopically in order to follow what is known to happen in the superior colliculus and the dorsal pathways of the human brain[5], meaning that they each relate to a single visual sensor and respective image-plane. However, as discussed earlier, attention is multisensory and joint attention is a dynamic and evolving form of basic joint action; this means that saliency maps must subsequently be subjected to a multitude of different, complex, 3D-based transformations[6]:

- saliency maps resulting from visual and auditory perception must be both related and integrated into a single egocentric reference;
- saliency maps must be temporally registered in order to accommodate egomotion (i.e., head and body motion);
- the information conveyed by saliency maps must eventually be related in sensorimotor fashion, so as to perform all types of motor commands needed to perform joint attention (e.g., head turns, gaze shifts, pointing gestures, etc).

Recent research efforts have introduced spherical representations of egocentric space to deal with these issues. For instance, Schillaci, Bodiroža, and Hafner [102] proposed a saliency-based attentional model combined with a tesselated egosphere, where information about salient areas is stored in the egosphere's edges, as proposed by Peters, Hambuchen, Kawamura, and Wilkes [118] and Fleming, Peters, and Bodenheimer [119]. In another example, Ferreira, Dias et al. [120, 121] proposed a log-spherical, egocentric inference grid used in the context of human-robot interaction to implement visuoauditory entropy-based exploration and also saliency-driven attentional behaviours [122], shown to be both suitable for real-time implementation [123] and for learning attentional sets for top-down modulation [124].

So, how does one close the action-perception loop in an egocentric perspective? Whilst active perception has been mostly approached in the perspective of devising solutions for choosing the appropriate focus of attention, the actual directing of one's senses towards locations of interest and the implications of the motor control associated to these actions have often been taken for granted, and recurrently understated. When considering the action-perception loop, research efforts in robotics have usually been focussed in determining the influence that predicting future perceptions has on the current choice of action: this, in fact, relates to the classical problem of modelling closed-loop control. Contrastingly, closing the loop by predicting the influence of motor knowledge on perception has been much less investigated – notable exceptions to this rule can be found, for example, in the works of Gilet et al. [125], Chella and Macaluso [126]. However, evidence from behavioural studies in humans demonstrate that actions are planned and controlled in terms of their effect, and that this accounts for much of our success as perceptual beings [127]. In fact, it has been shown that humans simulate the effects of their actions in advance using what is called *efference* or *efferent copies* [128]. Motor control in active perception complicates even further when considering head-eye coordination, including vergence for fixation on close objects.

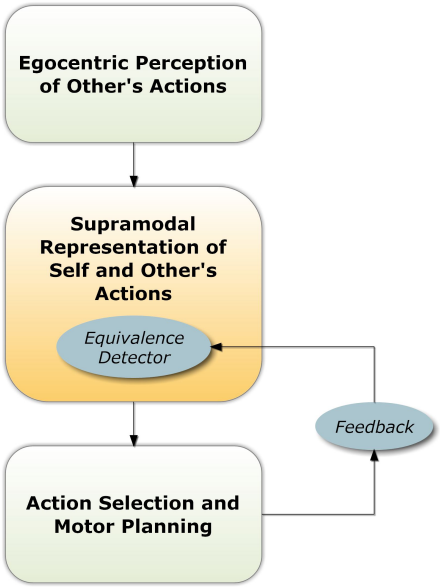

Figure 6. The active intermodal mapping (AIM) hypothesis of Meltzoff and Moore [129] adapted to the general context of joint attention.

The active intermodal mapping (AIM) hypothesis of facial imitation by Meltzoff and Moore [129] posits that infants match observations of adults with their own proprioception using a modality-independent representation of state. Mismatch detection between infant and adult states is performed in this modality-independent space. Consequently, infant actions and corresponding motor commands cause proprioceptive feedback, closing the action-perception loop – see Fig. 6. Note that this hypothesis is, in a way, the dyadic equivalent of the efferent copy concept – simulation, in this case, is not of one's own action, but of the other's. The AIM hypothesis has at times been reproduced in the context of joint attention, as it clearly indicates a roadmap to implementing the ability of relating the egocentric reference with the interlocutor's reference, and solutions based on its modality-independent representation have arguably become the most popular intermediaries from egocentric to dyadic representations. Nevertheless, attention detection mechanisms precede such a skill and therefore must be an integral part of a cognitive framework for joint attention.

Many machine vision systems have looked at the problems of identifying cues that indicate attention in order to establish dyadic representations, such as deictic pointing [130], head

[5]The dorsal pathway, and also the phylogenetically preceding superior colliculus (shared by all vertebrates, from mammals to more primitive creatures such as reptiles, conserved through evolution due to its fundamental role in survival [116]), both mediate attention and the perceptual control of action. For more information regarding these brain sites and for a summary on their specific roles in visual attention, please refer to [117] and [24, 26], respectively.

[6]As is also known to happen in the human brain.

Table II
COMPARISON OF SURVEYED RESEARCH WORK IN TERMS OF THE USE OF ATTENTION DETECTION AND DYADIC AND TRIADIC MECHANISMS AND REPRESENTATIONS.

Study	Attention Detection			Representation[a]
	Head Pose/Gaze Detection	Pointing Detection	Utterance Detection	
Scassellati [46, 48]	+	+	-	RSM
Breazeal and Scassellati [49–52]	-	-	+	RSM
Breazeal [54–56]	-	+	+	AIM (RSM + PG + MS) + EO3D
Doniec and Scassellati [57]	+	-	-	RSM + EO3D
Ishiguro et al. [59–61]	-	-	+	?
Nagai et al. [62–65]	-[b]	-	-	RSM + SOM
Nagai et al. [66]	-[b]	-	-	RSM + SOM
Nagai et al. [67]	-	-	-	RSM
Sumioka et al. [68]	-[b]	-	-	RSM + SOM
Hafner and Kaplan [74]	-	+	-	RSM + MLP
Kozima and Yano [77, 78]	+	-	-	-
Haasch et al. [79]	-	+	+	OSM
Hoffman et al. [81, 82]	+	-	-	AIM (RSM + CM + MS)
Yücel et al. [83]	+	-	-	RSM + EO3D
Yücel et al. [84]	+	-	-	RSM + EO3D
Andry et al. [88]	-	-	-	RSM + MS
Boucenna et al. [89]	+	-	-	RSM + MS
Ito and Tani [91–93]	-	-	-	RSM + MS
Romero et al. [95–97]	+	-	+	RSM
Huang and Thomaz [99, 100]	-	-[c]	+	RSM

In general: + implemented; - not implemented; ? not clear from author's description.

[a] RSM – retinotopic feature or saliency map; EO3D – 3D egocentric object position detection; SOM – self-organising map of other's face pattern for sensorimotor mapping; MLP - multi-layer-perceptron coding left or right pointing by other; OSM – "object scene model", consisting of a list of objects related to 3D positions and respective attributes; AIM – active intermodal mapping [129] adaptation; MS – motor/effector space; PG - "posegraph" referred to motor/effector space; CM - correspondence model in egocentric reference.

[b] In this case, although head pose and gaze direction are not directly measured, gaze following is learned indirectly through self-other sensorimotor mapping.

[c] In this case, a paper pointer with an ARToolKit marker was used.

pose [131], or gaze direction [132] in vision, or utterance detection and speech recognition [133] in audition, impelled by the fact that in the past few years it has become practical to use these systems in real-time robotic systems [32]. Note that, although the generation of most of these cues, such as pointing, for example, is clearly a voluntary action, *attention detection based on these cues is mostly an involuntary response*[7]. In fact, systems that use a conjunction of stimulus-driven attention mechanisms and deictic pointing gesture detection in HRI have been a recent focus of research – see, for example, the work by Schauerte, Richarz, and Fink [134]. Unfortunately, most state-of-the-art solutions are still far from producing results remotely close to human performance. For example, most of the algorithms used for head pose and gaze direction estimation are very dependent on frontal views of the interlocutor's face and/or a full set of facial features to be able to perform satisfactorily [131, 132].

The models presented in section III are compared in Table II in terms of their internal representations of perception and action and use of attention detection cues.

As can be seen in this table, attention detection processes specifically designed for joint attention have been the object of a substantial amount of research – 70% of the surveyed approaches use some sort of attention detection method to drive joint attention. In most other cases, even if head pose and gaze direction are not directly measured, gaze following is learned indirectly through self-other sensorimotor mapping. For instance, head pose and gaze detection has been used in more than half of the surveyed approaches (55%), either through preprogrammed modelling as an innate ability (40%), or resulting from a learning process (15%), in which case the actual learning becomes one of the main research objectives. Whenever auditory sensing is used (35% of the surveyed work, according to Table I), utterance detection, predictably, is nearly always used, with one exception (in other words, in approximately 86% of the 7 auditory-capable systems). Deictic pointing is the cue less used for detecting attention (20% of the overall sample), mainly because it is hard to model – see the work of Rouanet, Oudeyer et al. [135–137] for a discussion on this matter and possible alternatives; also, note how Huang and Thomaz [99, 100] circumvent this issue. A very small minority (15%) represent exceptions that do not use attention detection at all: instead, the authors rely on the robot's own attention processes (saliency, motion detection, etc.) to somehow indirectly make the robot attend to the same object as its interlocutor (a phenomenon referred to by Kaplan

[7] Magicians have used these involuntary responses to manipulate the attention of their audiences to their advantage for centuries.

and Hafner [29, 30] as "coincidental simultaneous looking").

On the other hand, representations for solving the egocentric-dyadic-triadic prediction problem in the work reviewed in this text, as hinted earlier, are mainly dependent on retinotopically-arranged feature or saliency maps (RSM – 85%). As a matter of fact, quite a few use these maps directly to drive attention, without resorting to more sophisticated motor mapping (approximately 29% of the 17 approaches that use RSM, 25% overall). Conversely, only two approaches (10%) aim for a more complete solution by modelling the "simulation of other" through an adaptation of the AIM hypothesis.

Again, it is our assessment that the overall complexity and tractability of modelling joint attention, together with the uncertainty inherent to this type of interaction, has limited the comprehensiveness of current approaches.

C. Goal, Action and Behaviour Selection – Regulating and Learning Social Interaction

A robot is incapable of engaging in social interaction if it does not have an inkling of what socially-relevant task it is carrying out, and how and when to switch between behaviours that might be essential to accomplish the task at hand. On the other hand, at first glance the nature and number of possible socially-relevant tasks seems intractable for current technologies in artificial cognitive systems.

To circumvent this problem, part of the research work presented in section III is based on the definition of small sets of reasonable basic social behaviours, goals and respective tasks, and simple heuristics to switch between the latter, mostly to validate the capabilities of social engagement of the proposed robots, usually with some grounding in developmental studies in humans.

On the other hand, emotion recognition (analysis) [138], and, to a lesser extent, emotion emulation (synthesis) [5] have been a subject of a great deal of recent research. In the context of joint attention, a considerable amount of work includes emotion recognition to infer acknowledgement and validation of correct RJA from the expressions and demeanour of the other, while emotion emulation has also been used to build expressive robots that convey analogous emotional content to their interlocutor.

Finally, turn-taking is a pervasive process in social interaction: turn-taking occurs during any form of communication, where the interlocutors need to take turns in order to successfully understand one another. In fact, turn-taking can be seen as two or more agents sharing a resource that cannot be allocated to more than one agent at any given time [139]. Consequently, even when dealing with primal behaviours such as joint attention, where turn-taking corresponds to the RJA-(EJA)-IJA cycle, tackling this problem is paramount, and mostly an open research question [29], although some principled work has already been carried out to study turn-taking in other, very specific scenarios, such as spoken dialogue generation between robots and humans – see, for example, Raux and Eskenazi [140].

As explained earlier, all of these processes modulate automatic attentional mechanisms as top-down influences, and therefore they are absolutely relevant to this survey. Consequently, the models presented in section III are compared in Table III in terms of:

- the potential top-down influences, such as which phases of joint attention are modelled (given that they represent, in fact, subtasks within the context of this social interaction);
- a qualification of the set of emulated goals;
- the emotions and behaviours that may be driving the robot;
- whether turn-taking is specifically taken into account.

The assessment of Table III yields the same observations as in the previous discussions: potential top-down influences are scarce (35%), robots with the ability to either infer or expressively instil emotion even less (25%), while the clear majority of the surveyed work only deal with a single phase of the regulation of joint attention (65%).

D. Experimental Validation Methods and Criteria

Although not fundamental for our review, an interesting and important matter to investigate are the methods and criteria used by the teams involved in the research surveyed in this text to tackle the issue of validating and benchmarking their systems, given how challenging it is to conceptualise the nature of the underlying processes and dynamics of joint attention.

Scassellati [48] suggested that techniques to evaluate joint attention mechanisms such as the Autism Diagnostic Interview and the Autism Diagnostic Observation Schedule [141] – used, for instance, in the diagnosis and assessment of autism and related disorders in children and adults[8] – could be adapted to validate robotic implementations with only minimal modifications. However, in general, research work in robotic implementations of joint attention have applied much less elaborate evaluation methods and criteria.

The most ubiquitous quantitative criterion for joint attention in robotics seems to be the *detection success rate measure* (or its counterpart, *total detection error*), applied in experiments where the robot's focus of attention is matched to its interlocutor's, by comparing the robotic observer's expectation to the other's deictic pointing or gazing targets – examples of work using this evaluation method would be [56, 57, 68–70, 75, 81, 89, 91, 99]. Some of these approaches use multiple conditions, such as varying object positions, or, less frequently, lighting conditions (e.g., [66]) or interaction sophistication (e.g., [99], in which the effect of presence or absence of robot RJA is studied). In some cases, the temporal or trial-by-trial evolution of this success rate is measured (e.g., [81, 91, 99].

Other examples of quantitative criteria would be *time-to-error detection* and *length of interaction by time or number of utterances* [56], human reaction time in RJA (also called "social delay") [57], and evolution of robot reaction time in RJA [81]. Time to learning convergence is also important when learning is used (e.g., [64, 66]). Another interesting (albeit controversial) quantitative measure of joint attention success

[8]See the work by Dautenhahn and Werry [142] for a discussion on the use of interactive robots in autism therapy.

Table III
COMPARISON OF SURVEYED RESEARCH WORK IN TERMS OF THE USE OF POTENTIAL TOP-DOWN INFLUENCES.

		Emotions		Joint Attention Regulation	
Study	Emotion/Drive/Action Sets[a]	Analysis	Synthesis	Interaction Cycle Phases[b]	Explicit Turn-Taking
Scassellati [46, 48]	CA + AO + CJA	-	-	(RJA + IJA)[c]	-
Breazeal and Scassellati [49–52]	BE + BD + SL + S + P	+	+	IJA + RJA	-
Breazeal [54–56]	BE + BD + ...	+	+	IJA + RJA	-
Doniec and Scassellati [57]	CA + AO + CJA	-	-	IJA (LP) + RJA	-
Ishiguro et al. [59–61]	-	-	+	IJA	-
Nagai et al. [62–65]	-	-	-	RJA (LEC + LGF)	-
Nagai et al. [66]	-	-	-	RJA (LEC + LGF)	-
Nagai et al. [67]	-	-	-	IJA + RJA + EJA	-
Sumioka et al. [68]	-	-	-	RJA (LEC + LGF)	+
Hafner and Kaplan [74]	-	-	-	RJA	-
Kozima and Yano [77, 78]	...[c]	-	-	RJA + EJA	-
Haasch et al. [79]	-	-	-	RJA	-
Hoffman et al. [81, 82]	-	-	-	RJA	-
Yücel et al. [83]	-	-	-	RJA (EC + GF)	-
Yücel et al. [84]	-	-	-	RJA (EC + GF)	-
Andry et al. [88]	-	-	-	-	+
Boucenna et al. [89]	-	+	+	RJA (LEC + LGF)	-
Ito and Tani [91–93]	-	-	-	-	+
Romero et al. [95–97]	CA + S + P + ...	-	+	RJA (LEC + LGF)	-
Huang and Thomaz [99, 100]	CA + AO + CJA	-	-	RJA + EJA + IJA	+

In general: + implemented; - not implemented; ? not clear from author's description.

[a] CA – capture attention; AO – address object; CJA – check joint attention; SL – sleep; S – socialise; P – play; BE – basic emotions (anger + disgust + joy + sorrow + surprise/fear); BD – basic drives (social drive + stimulation drive + fatigue drive); "..." (ellipsis) – expandable action/task set.
[b] RJA – respond to joint attention; IJA – initiate joint attention; EJA – ensure/acknowledge joint attention; (L)EC – (learnt) maintaining eye contact; (L)GF – (learnt) gaze following; (L)P – (learnt) diectic pointing.
[c] Rudimentary, incomplete implementations.

is obtained by using the result of the robot's self-evaluation (for example, the evolution of the "joy" emotion in [89]).

In addition to quantitative criteria, some researchers have resorted to presenting questionnaires to human participants in the experiments, partially relying on their (hopefully unbiased) subjective assessment of success [56, 67, 99].

We suggest that ideas from the Feature Integration Theory of attention by Treisman and Gelade [143] could additionally be used in the extended context of joint attention in *feature and conjunction search experiments*[9] – in both cases a target is placed among distractors, and the objective is to successfully find the target and as quickly as possible [24]. In a first set of feature search experiments, a unique (multisensory) feature of the target pops-out comparing to distractors, and, in a second first set of conjunction search experiments, a combination of features is used. After measuring *reaction times* (RTs) and detection success or accuracy (DA) in non-guided experiments, a third set of experiments is conducted using joint attention. Reaction times will be typically fast for the first set and slower for the second set. After measuring RTs and DA in the non-guided experiments, the third set of joint attention experiments is conducted – RT and DA can then be benchmarked to assess the robot's RJA, under different experimental conditions.

[9] As far as the authors know, this has not been done before.

V. CONCLUSION

With this survey we have attempted to provide a deeper insight in the modelling and implementation of automatic multisensory attentional mechanisms for socially interactive robots, using joint attention, acknowledged as a basic joint action, as a most appropriate backdrop. We summarised the contributions already made in these matters in robotic cognitive systems research and analysed scientific issues which, in our point of view, remain to be adequately addressed.

In our opinion, socially interactive robots would greatly benefit from the development of probabilistic real-time frameworks implementing automatic attentional mechanisms. These would, in our opinion, effectively constitute middleware for the development of more intelligent and complete socially interactive robotic systems. In fact, we argue that this hierarchical/modular architecture-oriented approach would equip socially-skilled robots with fundamental innate and preacquired capabilities that humans already possess from a very early age; it is our belief that failing to do so will most probably ultimately result in robots exhibiting social deficits similar to those of individuals suffering from autism spectrum disorders.

However, we think that recent technological advances, such as the massively parallel implementations made possible by GPU computing and programmable logic devices, will allow

the development of probabilistic automatic attentional frameworks dealing with the challenges enumerated in section II-B. We therefore predict that, at the time that this text is being written, a new important phase in socially interactive robot development is about to begin. Therefore, we are positive that these advances will spur research that will undoubtedly tackle these issues with improved chances of success, eventually leading to the fulfilment of current robotics roadmaps for the years to come [144].

Additionally, we predict that advances in this respect will also collaterally influence many important fields, such as human-machine interaction or automatic surveillance systems, for which assessing intent or attention control might be crucial factors. On the other hand, since robots offer the possibility of studying the processes underlying joint attention in a repeatable and separable fashion, we also expect this research to shed further light on this important set of human social skills.

ACKNOWLEDGEMENT

We would like to thank our reviewers for their comments, corrections and suggestions, which improved immensely the quality of this text, and also the Editors of the IEEE Transactions on Autonomous Mental Development for their kind assistance.

This work was supported by the Portuguese Foundation for Science and Technology (FCT) and by the European Commission via the COMPETE programme [project grant number FCOMP-01-0124-FEDER-028914, FCT Ref. PTDC/EEI-AUT/3010/2012].

REFERENCES

[1] "Robot & Frank, Official Website," 2012. [Online]. Available: http://robotandfrank-film.com/

[2] IMDb, "Robot and Frank," 2012. [Online]. Available: http://www.imdb.com/title/tt1990314/

[3] C. Huijnen, A. Badii, H. v. d. Heuvel, P. Caleb-Solly, and D. Thiemert, "'Maybe It Becomes a Buddy, But Do Not Call It a Robot' – Seamless Cooperation between Companion Robotics and Smart Homes," in *Ambient Intelligence*, ser. Lecture Notes in Computer Science, D. V. Keyson, M. L. Maher, N. Streitz, A. Cheok, J. C. Augusto, R. Wichert, G. Englebienne, H. Aghajan, and B. J. A. Kröse, Eds. Springer Berlin Heidelberg, January 2011, no. 7040, pp. 324–329.

[4] A. Tapus, M. J. Mataric, and B. Scassellati, "Socially assistive robotics," *IEEE Robotics and Automation Magazine*, vol. 14, no. 1, p. 35, 2007.

[5] C. Rich and C. L. Sidner, "Robots and avatars as hosts, advisors, companions, and jesters," *AI Magazine*, vol. 30, no. 1, p. 29, 2009.

[6] K. Severinson-Eklundh, A. Green, and H. Hüttenrauch, "Social and collaborative aspects of interaction with a service robot," *Robotics and Autonomous Systems*, vol. 42, pp. 223–234, 2003.

[7] E. Guizzo and E. Ackerman, "The rise of the robot worker," *IEEE Spectrum*, vol. 49, no. 10, pp. 34–41, October 2012.

[8] J. Forlizzi and C. DiSalvo, "Service Robots in the Domestic Environment: A Study of the Roomba Vacuum in the Home," in *HRI '06 Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot interaction*, 2006, pp. 258–265.

[9] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, March 2003.

[10] I. Leite, C. Martinho, and A. Paiva, "Social Robots for Long-Term Interaction: A Survey," *International Journal of Social Robotics*, vol. 15, no. 4, pp. 291–308, 2013.

[11] "Social Robot," EC FP7, People Programme, Industry-Academia Partnerships and Pathways (IAPP), grant agreement 285870. [Online]. Available: http://paloma.isr.uc.pt/projects/socialrobot/

[12] "SPENCER (Social situation-aware perception and action for cognitive robots)," EC IP FP7 ICT 2011-9. [Online]. Available: http://www.spencer.eu/

[13] "ROBOTS@HOME," 2007–2010, EC STREP FP6-2005-IST-6. [Online]. Available: http://robots-at-home.acin.tuwien.ac.at

[14] "COGNIRON," 2004–2007, EC IP FP6-IST-002020. [Online]. Available: http://www.cogniron.org

[15] "KSERA (Knowledgeable SErvice Robots for Aging)," 2010–2013, EC STREP ICT FP7. [Online]. Available: http://ksera.ieis.tue.nl

[16] "CompanionAble: Integrated Cognitive Assistive & Domotic Companion Robotic Systems for Ability & Security," 2008–2012, EC ICT FP7. [Online]. Available: http://www.companionable.net

[17] "DEXMART – DEXterous and autonomous dual-arm/hand robotic manipulation with sMART sensory-motor skills: A bridge from natural to artificial cognition," 2008–2012, EC IP ICT-FP7 216239. [Online]. Available: http://www.dexmart.eu

[18] M. A. Goodrich and A. C. Schultz, "Human-Robot Interaction: A Survey," in *Human-Computer Interaction*, 2007, vol. 1, no. 3, pp. 203–275.

[19] C. Moore and P. J. Dunham, Eds., *Joint attention: Its origins and role in development*. Hillsdale, NJ, England: Lawrence Erlbaum Associates, Inc, 1995, vol. vii.

[20] J. Zlatev and C. Balkenius, "Introduction: Why "Epigenetic Robotics"?" in *Proceedings of the First International Workshop on Epigenetic Robotics*, vol. 85, 2001, pp. 1–4.

[21] L. Berthouze and G. Metta, "Epigenetic robotics: modelling cognitive development in robotic systems," *Cognitive Systems Research*, vol. 6, no. 3, p. 189–192, 2005.

[22] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini, "Developmental robotics: a survey," *Connection Science*, vol. 15, no. 4, pp. 151–190, 2003.

[23] O. Déniz, M. Hernández, J. Lorenzo, and M. Castrillón, "An engineering approach to sociable robots," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 19, no. 4, pp. 285–306, 2007.

[24] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 1, p. 6, 2010.

[25] A. Borji and L. Itti, "State-of-the-art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–22, 2012.

[26] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature Reviews Neuroscience*, vol. 3, pp. 201–215, March 2002.

[27] J. Aloimonos, I. Weiss, and A. Bandyopadhyay, "Active Vision," *International Journal of Computer Vision*, vol. 1, pp. 333–356, 1987.

[28] R. Bajcsy, "Active perception vs passive perception," in *Third IEEE Workshop on Computer Vision*, Bellair, Michigan, 1985, pp. 55–59.

[29] F. Kaplan and V. V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, no. 2, pp. 135–169, 2006.

[30] F. Kaplan and V. Hafner, "The challenges of joint attention," in *Proceedings of the 4th International Workshop on Epigenetic Robotics*. Genoa: Lund University Cognitive Science Studies, 2004, pp. 67–74.

[31] S.-J. Blakemore, P. Boyer, M. Pachot-Clouard, A. Meltzoff, C. Segebarth, and J. Decety, "The detection of contingency and animacy from simple animations in the human brain," *Cerebral Cortex*, vol. 13, no. 8, pp. 837–844, August 2003, PMID: 12853370.

[32] B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, pp. 13–24, 2002.

[33] M. Asada, K. Hosoda, Y. Kuniyoshi, H. Ishiguro, T. Inui, Y. Yoshikawa, M. Ogino, and C. Yoshida, "Cognitive developmental robotics: A survey," *IEEE Transactions on Autonomous Mental Development*, vol. 1, no. 1, pp. 12–34, 2009.

[34] M. Lungarella and G. Metta, "Beyond gazing, pointing, and reaching: A survey of developmental robotics," in *Proceedings of the 3rd International Workshop on Epigenetic Robotics*, 2003, pp. 81–89.

[35] M. Carpenter and K. Liebal, "Joint attention, communication, and knowing together in infancy," in *Joint Attention: New Developments in Psychology, Philosphy of Mind, and Social Neuroscience*. MIT Press, 2011, ch. 7, p. 159–182.

[36] A. Fiebich and S. Gallagher, "Joint attention in joint action," *Philosophical Psychology*, vol. ahead of print, pp. 1–17, 2012.

[37] J. Bullier, "Integrated model of visual processing," *Brain Research Reviews*, vol. 36, pp. 96–107, 2001, review.

[38] D. Kahneman, *Thinking, Fast and Slow*. Farrar, Straus and Giroux, October 2011.

[39] M. Corbetta, G. Patel, and G. L. Shulman, "The Reorienting System of the Human Brain: From Environment to Theory of Mind," *Neuron*, vol. 58, no. 3, pp. 306–324, May 2008.

[40] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, pp. 219–227, 1985.

[41] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research*, vol. 40, pp. 1489–1506, 2000.

[42] J. J. Campos and C. Stenberg, "Perception, appraisal, and emotion: The onset of social referencing," in *Infant social cognition: Empirical and theoretical considerations*, M. E. Lamb and L. R. Sherrod, Eds. Hillsdale, NJ: Erlbaum, 1981, vol. 273-314, p. 314.

[43] C. Spence and V. Santangelo, "Auditory attention," in *Oxford Handbook of Auditory Science*, C. Plack, Ed. Oxford: University Press, 2010, ch. 11, pp. 249–270.

[44] H. Johansen-Berg and D. M. Lloyd, "The Physiology and Psychology of Selective Attention to Touch," *Frontiers in Bioscience*, vol. 5, pp. d894–904, 2000.

[45] Y. Sakaguchi and K. Nakano, "Haptic recognition system with sensory integration and attentional perception," in *Multisensor Fusion and Integration for Intelligent Systems, 1994. IEEE International Conference on MFI'94.*, 1994, pp. 288–295.

[46] B. Scassellati, "Imitation and mechanisms of shared attention: A developmental structure for building social skills," in *Proc. Autonomous Agents 1998 workshop Agents in Interaction-Acquiring Competence through Imitation, Minneapolis, MO (August 1998)*, 1998.

[47] R. A. Brooks, C. Breazeal, M. Marjanovic, B. Scassellati, and M. M. Williamson, "The Cog project: Building a humanoid robot," in *Computation for metaphors, analogy, and agents*. Springer, 1999, p. 52–87.

[48] B. Scassellati, "Investigating models of social development using a humanoid robot," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 4, 2003, p. 2704–2709.

[49] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," in *Proc. Int. Joint Conf. on Artificial Intelligence*, 1999, pp. 1146–1151.

[50] ——, "Infant-like social interactions between a robot and a human caregiver," *Adaptive Behavior*, vol. 8, no. 1, pp. 49–74, 2000.

[51] C. Breazeal, A. Edsinger, P. Fitzpatrick, and B. Scassellati, "Active Vision for Sociable Robots," *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 31, no. 5, pp. 443–453, September 2001.

[52] C. Breazeal, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 119–155, 2003.

[53] C. Breazeal and B. Scassellati, "Robots that imitate humans," *Trends in Cognitive Sciences*, vol. 6, pp. 481–487, 2002.

[54] C. Breazeal, G. Hoffman, and A. Lockerd, "Teaching and working with robots as a collaboration," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 3, 2004, pp. 1030–1037.

[55] C. Breazeal, D. Buchsbaum, J. Gray, D. Gatenby, and B. Blumberg, "Learning from and about others: Towards

using imitation to bootstrap the social understanding of others by robots," *Artificial Life*, vol. 11, no. 1-2, pp. 31–62, 2005.

[56] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, 2005 (IROS 2005)*, 2005, p. 708–713.

[57] M. W. Doniec, G. Sun, and B. Scassellati, "Active learning of joint attention," in *6th IEEE-RAS International Conference on Humanoid Robots*, 2006, pp. 34–39.

[58] M. Asada, K. F. MacDorman, H. Ishiguro, and Y. Kuniyoshi, "Cognitive developmental robotics as a new paradigm for the design of humanoid robots," *Robotics and Autonomous Systems*, vol. 37, no. 2-3, p. 185–193, 2001.

[59] H. Ishiguro, T. Ono, M. Imai, T. Maeda, T. Kanda, and R. Nakatsu, "Robovie: an interactive humanoid robot," *Industrial robot: An international journal*, vol. 28, no. 6, pp. 498–504, 2001.

[60] H. Ishiguro, T. Kanda, K. Kimoto, and T. Ishida, "A robot architecture based on situated modules," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'99)*, vol. 3, 1999, p. 1617–1624.

[61] M. Imai, T. Ono, and H. Ishiguro, "Physical relation and expression: Joint attention for human-robot interaction," *IEEE Transactions on Industrial Electronics*, vol. 50, no. 4, pp. 636–643, 2003.

[62] Y. Nagai, M. Asada, and K. Hosoda, "A developmental approach accelerates learning of joint attention," in *Proceedings of the 2nd International Conference on Development and Learning*, 2002, pp. 277–282.

[63] Y. Nagai, K. Hosoda, A. Morita, and M. Asada, "A constructive model for the development of joint attention," *Connection Science*, vol. 15, no. 4, pp. 211–229, 2003.

[64] Y. Nagai, "Understanding the Development of Joint Attention from a Viewpoint of Cognitive Developmental Robotics," Ph.D. dissertation, Osaka University, 2004.

[65] Y. Nagai, M. Asada, and K. Hosoda, "Learning for joint attention helped by functional development," *Advanced Robotics*, vol. 20, no. 10, pp. 1165–1181, 2006.

[66] Y. Nagai, "The Role of Motion Information in Learning Human-Robot Joint Attention," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, April 2005, pp. 2069–2074.

[67] Y. Nagai, A. Nakatani, and M. Asada, "How a robot's attention shapes the way people teach," in *Proceedings of the 10th International Conference on Epigenetic Robotics*, 2010, pp. 81–88.

[68] H. Sumioka, K. Hosoda, Y. Yoshikawa, and M. Asada, "Acquisition of joint attention through natural interaction utilizing motion cues," *Advanced Robotics*, vol. 21, no. 9, p. 983–999, 2007.

[69] H. Sumioka, M. Asada, and Y. Yoshikawa, "Causality detected by transfer entropy leads acquisition of joint attention," in *IEEE 6th International Conference on Development and Learning (ICDL 2007)*, July 2007, pp. 264–269.

[70] H. Sumioka, Y. Yoshikawa, and M. Asada, "Development of joint attention related actions based on reproducing interaction contingency," in *7th IEEE International Conference on Development and Learning (ICDL 2008)*, 2008, pp. 256–261.

[71] ——, "Learning of Joint Attention from Detecting Causality Based on Transfer Entropy," *Journal of Robotics and Mechatronics*, vol. 20, no. 3, pp. 378–385, 2008.

[72] Y. Yoshikawa, T. Nakano, M. Asada, and H. Ishiguro, "Multimodal joint attention through cross facilitative learning based on μx principle," in *7th IEEE International Conference on Development and Learning (ICDL 2008)*, 2008, pp. 226–231.

[73] H. Sumioka, Y. Yoshikawa, and M. Asada, "Reproducing interaction contingency toward open-ended development of social actions: Case study on joint attention," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 1, pp. 40–50, 2010.

[74] V. V. Hafner and F. Kaplan, "Learning to interpret pointing gestures: experiments with four-legged autonomous robots," in *Biomimetic Neural Learning for Intelligent Robots*. Springer, 2005, pp. 225–234.

[75] ——, "Interpersonal Maps and the Body Correspondence Problem," in *Proceedings of the Third International Symposium on Imitation in Animals and Artifacts*, Hertfordshire, UK, 2005, pp. 48–53.

[76] F. Kaplan and P.-Y. Oudeyer, "The progress-drive hypothesis: an interpretation of early imitation," *Models and mechanims of imitation and social learning: Behavioural, social and communication dimensions*, pp. 361–377, 2007.

[77] H. Kozima and H. Yano, "A robot that learns to communicate with human caregivers," in *Proceedings of the First International Workshop on Epigenetic Robotics*, 2001, p. 47–52.

[78] H. Kozima, "Infanoid – A Babybot that Explores the Social Environment," in *Socially Intelligent Agents*, ser. Multiagent Systems, Artificial Societies, and Simulated Organizations, K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, Eds. Springer US, January 2002, no. 3, ch. 19, pp. 157–164.

[79] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer, "A multi-modal object attention system for a mobile robot," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*, August 2005, pp. 2712–2717.

[80] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON – The Bielefeld Robot Companion," in *Proc. Int. Workshop on Advances in Service Robotics*, E. Prassler, G. Lawitzky, P. Fiorini, and M. Hägele, Eds. Stuttgart, Germany: Fraunhofer IRB Verlag, 2004, pp. 27–32.

[81] M. W. Hoffman, D. B. Grimes, A. P. Shon, and R. P. Rao, "A probabilistic model of gaze imitation and

shared attention," *Neural Networks*, vol. 19, no. 3, pp. 299–310, 2006.

[82] A. P. Shon, D. B. Grimes, C. L. Baker, M. W. Hoffman, S. Zhou, and R. P. Rao, "Probabilistic gaze imitation and saliency learning in a robotic head," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, 2005, pp. 2865–2870.

[83] Z. Yücel, A. Salah, C. Meriçli, and T. Meriçli, "Joint visual attention modeling for naturally interacting robotic agents," in *24th International Symposium on Computer and Information Sciences (ISCIS 2009)*, September 2009, pp. 242–247.

[84] Z. Yücel, A. Salah, C. Meriçli, T. Meriçli, R. Valenti, and T. Gevers, "Joint Attention by Gaze Interpolation and Saliency," *IEEE Transactions on Cybernetics*, vol. 43, no. 3, pp. 829–842, 2013.

[85] Aldebaran Robotics, "Nao humanoid robot." [Online]. Available: http://www.aldebaran-robotics.com/eng/Nao.php

[86] FESTO, "Robotino robot platform." [Online]. Available: http://www.festo-didactic.com/int-en/learning-systems/education-and-research-robots-robotino/

[87] E. Carlson and J. Triesch, "A computational model of the emergence of gaze following," *Progress in Neural Processing*, vol. 15, pp. 105–114, 2004.

[88] P. Andry, P. Gaussier, S. Moga, J.-P. Banquet, and J. Nadel, "Learning and communication via imitation: An autonomous robot perspective," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 31, no. 5, pp. 431–442, 2001.

[89] S. Boucenna, P. Gaussier, and L. Hafemeister, "Development of joint attention and social referencing," in *2011 IEEE International Conference on Development and Learning (ICDL)*, vol. 2, August 2011, pp. 1 –6.

[90] T. Ikegami and H. Iizuka, "Joint attention and dynamics repertoire in coupled dynamical recognizers," in *In AISB 03: the Second International Symposium on Imitation in Animals and Artifacts*, 2003.

[91] M. Ito and J. Tani, "Joint attention between a humanoid robot and users in imitation game," in *Proc. of the Int. Conf. on Development and Learning (ICDL)*, 2004.

[92] ——, "On-line imitative interaction with a humanoid robot using a dynamic neural network model of a mirror system," *Adaptive Behavior*, vol. 12, no. 2, pp. 93–115, 2004.

[93] J. Tani and M. Ito, "Interacting with neurocognitive robots: A dynamical system view," in *Proceedings of the second international workshop on man-machine symbiotic systems*, Kyoto, Japan, 2005, pp. 123–134.

[94] M. Fujita, Y. Kuroki, and T. Ishida, "A small humanoid robot sdr-4x for entertainment applications," in *Proceedings of the 2003 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM 2003)*, vol. 2, 2003, pp. 938–943.

[95] C. A. Policastro, R. A. Romero, G. Zuliani, and E. Pizzolato, "Learning of shared attention in sociable robotics," *Journal of Algorithms*, vol. 64, no. 4, pp. 139–151, October 2009.

[96] R. da Silva, C. Policastro, and R. Romero, "Relational reinforcement learning applied to shared attention," in *International Joint Conference on Neural Networks (IJCNN 2009)*, June 2009, pp. 2943–2949.

[97] R. R. da Silva and R. A. F. Romero, "Modelling Shared Attention Through Relational Reinforcement Learning," *Journal of Intelligent & Robotic Systems*, vol. 66, no. 1-2, pp. 167–182, April 2012.

[98] K. Driessens, "Relational reinforcement learning," *AI Communications*, vol. 18, no. 1, pp. 71–73, 2005.

[99] C.-M. Huang and A. L. Thomaz, "Joint attention in human-robot interaction," in *Dialog with Robots: Papers from the AAAI Fall Symposium (FS-10-05)*. Association for the Advancement of Artificial Intelligence, 2010.

[100] ——, "Effects of responding to, initiating and ensuring joint attention in human-robot interaction," in *IEEE RO-MAN*, 2011, pp. 65–71.

[101] C. Muhl, Y. Nagai, and G. Sagerer, "On constructing a communicative space in HRI," in *KI 2007: Advances in Artificial Intelligence*. Springer, 2007, pp. 264–278.

[102] G. Schillaci, S. Bodiroža, and V. V. Hafner, "Evaluating the effect of saliency detection and attention manipulation in human-robot interaction," *International Journal of Social Robotics*, vol. 5, no. 1, pp. 139–152, January 2013.

[103] C. L. Folk and R. Remington, "Top-down modulation of preattentive processing: Testing the recovery account of contingent capture," *Visual Cognition*, vol. 14, no. 4-8, pp. 445–465, 2006.

[104] R. A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, vol. 2000, no. 7, pp. 17–42, 2003.

[105] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, pp. 1395–1407, 2006.

[106] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Advances in neural information processing systems*, vol. 18, p. 547, 2006.

[107] ——, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, May 2009, PMID: 18834898 PMCID: PMC2782645.

[108] N. Bruce and J. Tsotsos, "Saliency based on information maximization," *Advances in neural information processing systems*, vol. 18, p. 155, 2006.

[109] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, March 2009, PMID: 19757944.

[110] R. Palermo and G. Rhodes, "Are you always on my mind? A review of how face perception and attention interact," *Neuropsychologia*, vol. 45, pp. 75–92, 2007.

[111] N. Kirchner and A. Alempijevic, "A Robot Centric Perspective on the HRI Paradigm," *Journal of Human-Robot Interaction*, vol. 1, no. 2, pp. 135–157, 2012.

[112] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, p. 23–36, 2006.

[113] J. K. Tsotsos, S. M. Culhane, W. Y. Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, October 1995.

[114] M. V. Butz, O. Sigaud, and P. Gérard, *Anticipatory Behavior in Adaptive Learning Systems: Foundations, Theories, and Systems*. Springer-Verlag, 2003, pp. 86–109.

[115] E. Bicho, W. Erlhagen, E. Sousa, L. Louro, N. Hipolito, E. C. Silva, R. Silva, F. Ferreira, T. Machado, M. Hulstijn, Y. Maas, E. De Bruijn, R. H. Cuijpers, R. Newman-Norlund, H. Van Schie, R. G. J. Meulenbroek, and H. Bekkering, "The power of prediction: Robots that read intentions," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012, pp. 5458–5459.

[116] C. Maximino, "Evolutionary Changes in the Complexity of the Tectum of Nontetrapods: A Cladistic Approach," *PLoS ONE*, vol. 3, no. 10, p. e3582, October 2008.

[117] K. J. Murphy, D. P. Carey, and M. A. Goodale, "The Perception of Spatial Relations in a Patient with Visual Form Agnosia," *Cognitive Neuropsyshology*, vol. 15, no. 6/7/8, pp. 705–722, 1998.

[118] R. A. Peters, K. A. Hambuchen, K. Kawamura, and D. M. Wilkes, "The sensory ego-sphere as a short-term memory for humanoids," in *Proc. of the IEEE-RAS Int'l Conf. on Humanoid Robots*, 2001, pp. 22–24.

[119] K. Fleming, R. Peters, and R. E. Bodenheimer, "Image Mapping and Visual Attention on a Sensory Ego-Sphere," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2006, pp. 241–246.

[120] J. F. Ferreira, J. Lobo, P. Bessière, M. Castelo-Branco, and J. Dias, "A Bayesian Framework for Active Artificial Perception," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 43, no. 2, pp. 699–711, April 2013.

[121] J. F. Ferreira and J. Dias, "Hierarchical Log-Spherical Inference Grid – An Unconventional Approach to Robotic Perception and Action," in *UARACIN – Workshop on Unconventional Approaches to Robotics, Automation and Control Inspired by Nature, 2013 IEEE International Conference on Robotics and Automation (ICRA 2013)*, May 2013.

[122] J. F. Ferreira, M. Castelo-Branco, and J. Dias, "A hierarchical Bayesian framework for multimodal active perception," *Adaptive Behavior*, vol. 20, no. 3, pp. 172–190, June 2012.

[123] J. F. Ferreira, J. Lobo, and J. Dias, "Bayesian real-time perception algorithms on GPU — Real-time implementation of Bayesian models for multimodal perception using CUDA," *Journal of Real-Time Image Processing*, vol. 6, no. 3, pp. 171–186, September 2011.

[124] J. F. Ferreira, C. Tsiourti, and J. Dias, "Learning emergent behaviours for a hierarchical Bayesian framework for active robotic perception," *Cognitive Processing*, vol. 13, no. 1, pp. 155–159, July 2012.

[125] E. Gilet, J. Diard, and P. Bessière, "Bayesian Action–Perception Computational Model: Interaction of Production and Recognition of Cursive Letters," *PLoS ONE*, vol. 6, no. 6, p. e20387, June 2011.

[126] A. Chella and I. Macaluso, "The perception loop in CiceRobot, a museum guide robot," *Neurocomputing*, vol. 72, pp. 760–766, 2009.

[127] W. Prinz, "Perception and action planning," *European Journal of Cognitve Psychology*, vol. 9, no. 2, pp. 129–154, 1997.

[128] J. F. Poulet and B. Hedwig, "New insights into corollary discharges mediated by identified neural pathways," *TRENDS in Neurosciences*, vol. 30, no. 1, pp. 14–21, 2007.

[129] A. N. Meltzoff and M. K. Moore, "Explaining facial imitation: A theoretical model," *Early development and parenting*, vol. 6, no. 34, p. 179–192, 1997.

[130] T. Darrell and A. Pentland, "Active gesture recognition using learned visual attention," in *Advances in Neural Information Processing Systems (NIPS)*, D. T. et al., Ed. MIT Press, 1996.

[131] E. Chutorian and M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," *IEEE Transactions on Pattern Analyses and Machine Intelligence*, vol. 31, no. 4, pp. 607–629, 2009.

[132] D. Hansen and Q. Ji, "In the Eye of the Beholder: A Survey of Models for Eyes and Gaze," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[133] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[134] B. Schauerte, J. Richarz, and G. A. Fink, "Saliency-based identification and recognition of pointed-at objects," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 4638–4643.

[135] P. Rouanet, J. Bechu, and P.-Y. Oudeyer, "A comparison of three interfaces using handheld devices to intuitively drive and show objects to a social robot: the impact of underlying metaphors," in *18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2009)*, October 2009, pp. 1066 –1072.

[136] P. Rouanet, P.-Y. Oudeyer, and D. Filliat, "An integrated system for teaching new visually grounded words to a robot for non-expert users using a mobile device," in *9th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2009)*, December 2009, pp. 391 – 398.

[137] P. Rouanet, J. Béchu, and P.-Y. Oudeyer, "A survey of interfaces using handheld devices to intuitively drive and show objects to a social robot," in *IEEE International Conference on Robot and Human Interactive Communication*, 2009.

[138] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[139] P. A. Raffensperger, R. Y. Webb, P. J. Bones, and A. I. McInnes1, "A simple metric for turn-taking in emergent

communication," *Adaptive Behavior*, vol. 20, no. 2, pp. 104–116, 2012.

[140] A. Raux and M. Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, D. Schlangen and B. A. Hockey, Eds. Morristown, NJ: Association for Computational Linguistics, 2008, pp. 1–10.

[141] F. R. Volkmar, R. Paul, A. Klin, and D. Cohen, Eds., *Handbook of autism and pervasive developmental disorders*. John Wiley & Sons New Jersey, 2005.

[142] K. Dautenhahn and I. Werry, "Towards interactive robots in autism therapy: Background, motivation and challenges," *Pragmatics & Cognition*, vol. 12, no. 1, p. 1–35, 2004.

[143] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[144] EURON, "EURON Research Roadmap V4.1," 2008.

João Filipe Ferreira (M'12) was born in Coimbra, Portugal, in 1973. He received the B.Sc. (five-year course), M.Sc., and Ph.D. degrees in electrical engineering and computers from the University of Coimbra, Coimbra, in 2000, 2005, and 2011, respectively.

He has been an Invited Assistant Professor at the University of Coimbra since 2011. He has also been a researcher at the Institute of Systems and Robotics (ISR), University of Coimbra, since 1999 (integrated member since 2011). His main research interests are spread out through three broad scientific themes: Artificial Cognition, Probabilistic Modelling and Autonomous Systems. He is currently the Principal Investigator (PI) for the CASIR project (Coordinated Control of Stimulus-Driven and Goal-Directed Multisensory Attention Within the Context of Social Interaction with Robots – PTDC/EEI-AUT/3010/2012), funded by the Portuguese Foundation for Science and Technology (FCT) and the European Commission via the COMPETE programme.

Dr. Ferreira is a member of the IEEE Robotics and Automation Society, the IEEE Systems, Man, and Cybernetics Society, and the IEEE Life Sciences Community.

Jorge Dias (M'96–SM'10) received the Ph.D. degree in electrical engineering with specialisation in control and instrumentation from the University of Coimbra, Coimbra, Portugal, in 1994.

He holds his research activities at the Institute of Systems and Robotics (ISR), University of Coimbra, and also at the Khalifa University of Science, Technology and Research, Abu Dhabi, UAE. His current research areas are computer vision and robotics, with activities and contributions in these fields since 1984. He has been the main researcher in several projects financed by the European Commission (Framework Programmes 6 and 7) and by the Portuguese Foundation for Science and Technology (FCT).

Dr. Dias is currently the officer in charge for the Portuguese Chapter for the IEEE Robotics and Automation Society and also the Vice President of "Sociedade Portuguesa de Robótica".