CNRS Research Project

Socially-Driven Autonomous Robots

for

Real-world Human-Robot Interactions

Pr. Séverin Lemaignan

Proposed host laboratory:
Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS-CNRS)

Summary

AI and robots are increasingly part of our everyday lives, eg supporting our ageing society or assisting teachers in classrooms. In this context, how to ensure *by design* that these social robots have a positive social impact? This question is the backbone of my research project, and my specific objective is to create within 5 to 10 years socially-intelligent and responsible robots, that (1) will have recognised social utility, and (2) will see long-term acceptance by their users.

I formulate two main hypotheses: (1) this objective can only be achieved if the robot is socially-driven: the robot's behaviours must be intrinsically driven by the intention to support positive human-human interactions (a *social teleology*). How this general principle translates into specific guidelines and algorithms – while taking into account the principles of a responsible AI – is a central research area of my project.

(2) Long-term acceptance requires genuine involvement of the end-users at every step of the design process. To this end, my project introduces a novel methodology involving 'public-in-the-loop' machine learning: the large scale participation of end-users, over extended periods of time, to teach robots how to become good and responsible social helpers.

My research programme explores these hypotheses with a scientifically ambitious and highly technical work programme. It includes basic research and conceptual framing; extensive, beyond-state-of-art, technical developments; and an ambitious experimental programme, centered on field deployments of social robots in public spaces.

This research programme is also important: it will provide us with a much better understanding and intellectual framing of what social robots *can* and *should* be, paving the way for their much broader adoption in the coming years: I will lead the design and implementation of socially-intelligent robots that are socially useful, acceptable in the long-term, and ethically responsible.

Contents

Long-term vision and ground-breaking nature of the project

This research project is about designing and delivering a ground-breaking embodied AI for socially intelligent robots, with long-term social utility and demonstrated acceptance in the real world.

This breakthrough is made possible by a combination of novel methodologies and the principled integration of complex socio-cognitive capabilities:

- **crowd-sourced social interaction patterns;**
- **'public-in-the-loop' machine learning;**
- **a novel spatio-temporal and social model of the robot's environment;**
- **novel, non-repetitive, social behaviour production based on generative neural networks;**
- **and finally, an integrative cognitive architecture, driven by long-term social goals.**

In addition, I will deliver the conceptual and ethical framework required to further support the public debate and policy making process around social robots, and concretely demonstrate lifescale applications of this technology with ambitious, long-term deployments of autonomous robots in high impact, social environments.

Closely aligned with the national and European research priorities, this research project creates a excellent opportunity to assert the CNRS and Europe as worldwide leaders in Social and Intelligent Robotics.

State of the art: real-world social robots and impact on the society

Social robotics is a disruptive field, with a profound impact on society and economy (Williams 2020). A recent report from the United Nations about the impact of the technological revolution on labour markets stated that AI and robotics are expected to radically change the labor market world-wide destroying some job categories and creating others (Bruckner et al. 2017). Social robotics, however, is still an young, emerging, research-active field. The expectations are high, in multiple application domains: elderly care, customer service (in airports and shopping malls, for instance), education, child development, and autonomous vehicles to name a few (Baillie et al. 2019). However, whereas both computer-based AI applications, and traditional industrial robots already have a significant economic impact, social robots have not reached that point yet. Significantly, the recent failures of several companies investing in social robotics, like Jibo, Kuri, Willow Garage and Anki, and the major setbacks of companies like SoftBank, who designed and deployed hundreds of Pepper robot in their shops, before renouncing a few months later due to the poor reception by the customers, show that these technologies are not yet mature (Tulli et al. 2019).

Indeed, understanding *why* these robots have failed, is one of the active debate within the Human-Robot Interaction community (Hoffman 2019), with only a handful of qualitative studies on this question (Dereshev et al. 2019; Graaf et al. 2017). Proposed explanations include the lack of perceived usefulness (robot seen purely as a toy); the limited liveliness of the robot that become rapidly predictable and repetitive (Lemaignan et al. 2014); the poor management of expectations, where user over-ascribe cognitive capabilities that do not match the reality. The community agrees however that the crux of the issue is achieving long-term social engagement (Yang et al. 2018; Hoffman 2019)

Research is however seemingly hitting a wall to further progress towards socially meaningful long-term interactions. For instance, in their large review of research in robotics for education, Belpaeme et al. (2018) point to the shortcomings that prevent further development of effective,

long-term social robotics in educative settings: the need for a correct interpretation of the social environment; the difficulty of action selection; the difficulty of pacing generated behaviours: three issues that underpin long-term engagement.

Attempts at long-term human-robot interactions are nevertheless becoming more common (Kunze et al. 2018; Leite et al. 2013), with a number of studies involving social robots deployed in real-world settings (for instance in schools (Leite et al. 2014; Westlund et al. 2017; Lemaignan et al. 2016; Coninx et al. 2016), homes (Graaf et al. 2017) and care centres (Hawes et al. 2017; Winkle et al. 2020a)) over relatively long periods of time (up to 2 or 3 months at a time). Even though these robots are typically not fully autonomous, they do exhibit a level of autonomy, either by handling autonomously a relatively broad range of shallow tasks (eg, a butler-like robot answering simple questions, like in Hawes et al. (2017) or in the H2020 MuMMer (Heikkilä et al. 2018) and FP7 Spencer project (Triebel et al. 2016)), or a narrow, well-specified complex task (for instance, supporting exercising in a gym, as I did in Winkle et al. (2020a)). However, general purpose, long-term interaction is still an open question.

Framing and objectives

The overall aim of my research programme is to **create, sustain and better understand the dynamics of responsible, long-term social human-robot interactions**. This translates into three overarching, long-term research questions:

- What are the public expectations with respect to the role of social robots, and how can we **collectively design** principles ensuring **autonomous**, yet **responsible, beneficial, socially acceptable robots**?
- What are the conceptual, algorithmic and technical prerequisites to design and implement such an autonomous & responsible robots? in particular, what AI is required to **sustain long-term engagement** between end-users and a robot?
- What new ethical questions are raised by long-term social interaction with an artificial agent, and in particular, how to balance **autonomy** of the robot with **behaviour transparency** and **human oversight**?

From these questions, I derive the following five objectives that are the guiding principles of my research programme:

O1: conceptual framing To construct a solid conceptual framing around the multidisciplinary question of responsible human-robot interactions, answering questions like: What should motivate the robot to step in and attempt to help? or: What social norms are applicable to the robot behaviours? I will investigate the basic principles of responsible social interactions, that must form the foundations of a socially useful robot, accepted and used in the long run. Using user-centred design and participatory design methodologies, I will identify the determinants and parameters of a responsible social intervention, performed by a socially-driven robot, and formalise them in practical principles.

O2: real-time social modeling Critical to long-term sustain engagement is the understanding by the robot of its social environment: this objective is to create the novel cognitive capability of artificial *social situation assessment* and enable the robot to represent real-time social dynamics in its environment, I will significantly extend and integrate the current state-of-art in social signal processing and spatio-temporal modeling (so-called *situation assessment*) with my recent research in social state modeling, exploring as well novel avenues like *social embeddings* to enable better machine learning.

O3: congruent social behaviours production I want to explore and create novel ways of producing non-repetitive, socially-congruent, expressive behaviours. This includes for instance expressive social motions using state-of-the-art generative neural networks to co-invent with artists (eg choreographers) a novel 'body language' for robots.

O4: embodied AI breakthrough I aim to create robot behaviours that are perceived as purposeful and intentional (long-term goals), while being shaped by a user-created and user-controlled action policy. I will integrate long-term social goals, arising from the interaction principles of **O1**, with the social modeling capability of **O2** and the behaviours production of **O3** into a principled, goal-driven cognitive architecture. The breakthrough will come from combining these long-term social goals with bottom-up action policies, designed and learnt from the end-users using human-in-the-loop reinforcement learning.

I want to specifically test the following two hypotheses: first, that long-term social goals, if suitably co-designed with the public and stakeholders and properly integrated into the robot as a *social teleology*, will create the perception that the robot is intentional and purposeful. This will in turn elicit sustained engagement from its human users.

Second, that human-in-the-loop machine learning can be used to ensure an additional layer of human oversight and a level of behavioural transparency. Human-in-the-loop reinforcement learning – as implemented in the SPARC approach that I have developed with my students and already used in complex social environments (Senft et al. 2017; Senft et al. 2019; Winkle et al. 2020b) – relies on an end-user 'teacher'. This teacher initially fully controls the robot (via teleoperation) while it learns the action policy, and then progressively relinquishes control up to a point where the robot is effectively autonomous. As I previsouly argued in Senft et al. (2019), this approach leads to increased control and ownership of the system, and as a result, increased trust from the end-users.

This objective also raises one additional question: how to *arbitrate* between a top-down action policy arising from the long-term goals and the bottom-up action policy learnt from the end-users? This question leads to objective **O4'**: To design a policy arbitration mechanism that preserves the robot's long-term intentional behaviour while effectively guaranteeing human control, ownership and oversight.

O5: ambitious field research Finally, the last major objective of my research project is to demonstrate the effectiveness of my approach in complex, real-world conditions. This means deploying the socially interactive robots in existing social eco-systems that are sufficiently complex and open to explore novel social interactions. My objective is also to show that this real-world deployment can be successfully driven by the 'end-to-end' involvement of all the end-users and stakeholders: from defining the robot's role, from the different perspective of each end-user, to actually designing and 'teaching' the robot what to do.

Together, these five objectives build a coherent and realistic pathway towards addressing the overall aim of my research programme: creating, sustaining and better understanding the dynamics of responsible long-term social human-robot interactions.

Research programme

These five scientific objectives underpin my research vision and scientific programme. This section presents how I intend to implement these objectives, i.e. what are the major research directions that I will research, develop and establish as active research fields in the coming years.

I intend to organise my research along **4+1 main research strands**:

- **Strand 1** focuses on advancing the **perception of complex social situations**, including modeling the complexity of humans and human group dynamics (objectives O1, O2);
- **Strand 2** investigates the **intelligent generation of social behaviours**, exploring novel techniques mixing immersive teleoperation and adversarial generative networks (objective O3);
- **Strand 3** aims at significantly progressing the state-of-art in **cognitive architectures** for robots, also accounting for and integrating end-users in the generation of cognitive behaviours (objectives O1, O4).
- **Strand 4** focuses on framing and practically advancing what responsible and safe AI means in the context of social robots. Critically, I propose a methodology enabling the co-construction of these guidelines with both the general public and ethics expert. This work will pave the way for an international framework and concrete guidelines for **responsible human-robot interactions** (objective O1).

Those four research strands are all underpinned by one additional research activity (objective O5), realised in the transversal **Strand 5**. Building on my own extensive experience in real-world deployments of autonomous social robots (as outlined in my academic track-record), I will establish an ambitious experimental programme, in close partnership with local institutions, based on the field's best practices that I have contributed to establish (Baxter et al. 2016).

Overview and coherence of the research programme

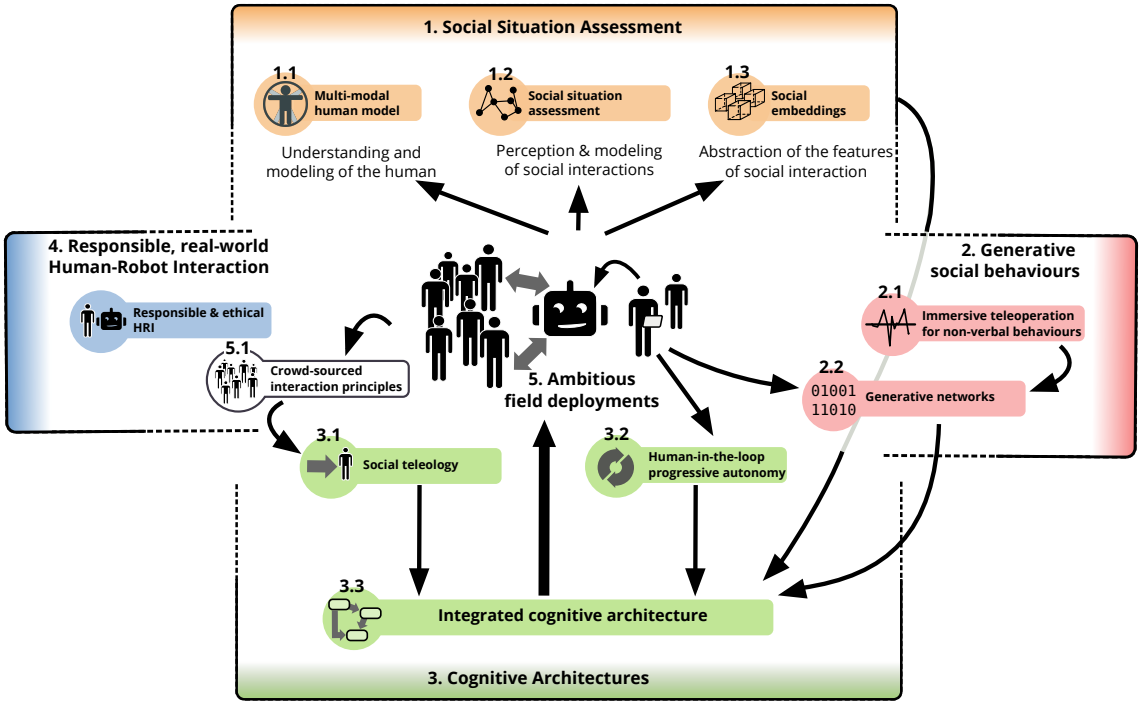

Figure 1: Overview of the research strands that I intend to develop as a CNRS research scientist.

These five strands are tightly coupled, and together will enable a major scientific and technical breakthrough in autonomous, socially-intelligent, robots.

Figure 1 gives an overview of how the research directions interact with each other. Fieldwork (**Strand 5**) plays a central role in my research programme, and appears in the centre of the figure. Indeed, the deployment of autonomous social robots in real-world, meaningful social spaces (eg. in schools, hospitals, museums and science centres, etc) will be integral to my research methodology, and will enable the development of 'public-in-the-loop' experiments (5.1): the public, by co-designing interventions, interacting and, at time, taking direct control of the robots, will shape what a useful and socially acceptable interaction looks like, and lead to the *definition of core interaction principles*. Using machine learning to learn from these field experiments (3.2), these core principles are in turn translated into algorithmic models, guiding the *social teleology* of the cognitive architecture (3.1).

The regular fieldwork I intend to conduct will also provide the source of data to feed into Strand 1: **Strand 1**, focusing on *social situation assessment*, researches, develops, and integrates all the components pertaining to the assessment of the spatio-temporal and social environment of the robot. Reference interaction situations and the interaction datasets required to support this research is directly drawn from the experimental fieldwork, as well as an additional, focused experimental programme on mental states modeling that I detail hereafter.

These perceptual capabilities are both continuously integrated into the robot's cognitive architecture (3.3), iteratively improving the socio-cognitive performances of the robot, and disseminated to the broader community through standardisation and integration to the international Robot Operating System (ROS) ecosystem.

Strand 2 looks into behaviour generation using immersive teleoperation to investigate novel non-verbal interaction modalities (2.1), combined with new developments in machine learning to learn and automatically generate them (2.2). In this research strand, I will focus on researching new way of automatically generating rich behaviours (including eg expressive gestures, expressive motions) that are non-repetitive and socially congruent. I intend to apply state-of-the-art deep generative networks to achieve this; as such, the research strand is data-intensive, and will use datasets acquired during the field deployments, as well as lab-recorded dataset of social interactions, using novel immersive techniques presented below. Similar to Strand 1, the newly developed capability of generating socially congruent behaviours is continuously integrated in the robot architecture.

Finally, an important part of my research programme contributes to the design and implementation of a novel, principled cognitive architecture for socially intelligent robots (**Strand 3**). In addition to the integration of Strand 1 and Strand 2, Strand 3 is also about researching and developing the socio-cognitive principles, or *drives*, of the architecture. They will be identified both from the 'public-in-the-loop' research and end-user engagement conducted in **Strand 4**, and an novel research on intrinsic social motivation that I detail below.

The following sections describe in greater detail each of these research strands, in the context of the current state-of-the-art.

Strand 1: Perception for robust real-world social situation assessment

My first research direction will look into integrating a full representation system for the social environment of the robot. It builds on existing state of art in *situation assessment* and *knowledge representation*, and extend it to the social sphere (Figure 2).

Indeed, knowledge representation and grounding is a fundamental building block for cogni-

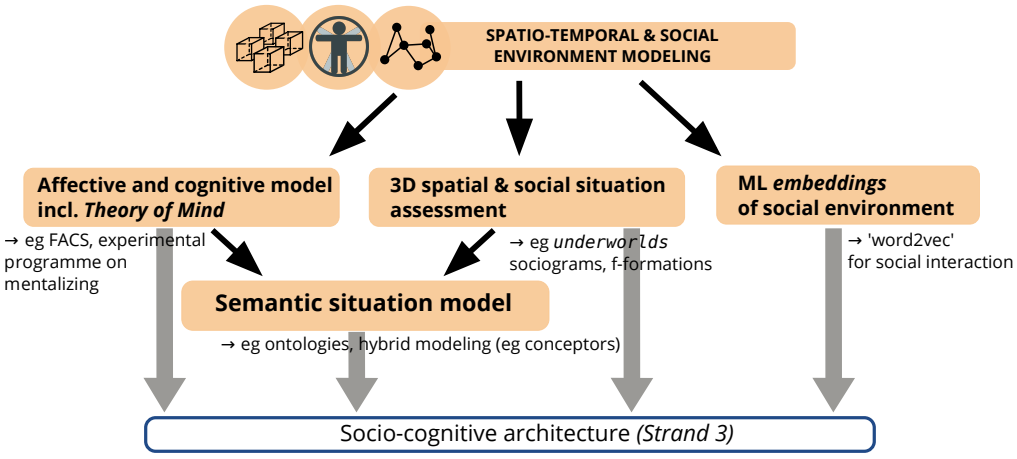

Figure 2: Social situation assessment, from feature extraction to 3D modelling, to social *embeddings* and symbolic reasoning. I will focus on using ROS for implementation, facilitating broad and lasting impact on the community.

tive architectures, as I have previously shown (Lemaignan et al. 2017). In fact this first research strand builds on existing work on symbolic knowledge representation (eg. (Tenorth and Beetz 2009; Lemaignan et al. 2010)) and my work on situation assessment (Lemaignan et al. 2018c; Sallami et al. 2019) to create a coherent system of representations for the cognitive architecture, that I will further improve with recent advances in symbolic (eg. data-driven semantic labelling, like the 4D convolution network MinkowskiNet (Choy et al. 2019)) and hybrid (like *conceptors* (Jaeger 2014)) representations capabilities.

The main contribution of the research strand, however, is on **augmenting this traditional spatio-temporal environment modeling with social representations**. I intend to work on three specific aspects: (1) modeling humans and social dynamics; (2) integrating these models into the cognitive perception of robots; (3) specifying and computing *social embeddings*, a low-dimensional synthetic representation of the social surroundings of the robots.

1.1 – Multi-modal human model; interaction and group dynamics

This first research activity focuses on the acquisition, processing and modelling of social signals (Gunes and Schüller 2017) to build a multi-modal model of the humans in the robot's vicinity. I have recently introduced a dataset of social interaction (Lemaignan et al. 2018b) that enables for the first time a quantitative, data-driven investigation of social dynamics. Promising initial results led me to uncover three latent constructs that underpin social interactions (Bartlett et al. 2019). I want to combine this emerging data-driven paradigm with my previous research on the psychosocial determinants of the interaction (anthropomorphism, trust, persuasion, etc.) to define and implement a holistic, multi-modal model of the human, suitable for real-time HRI.

From modelling the human, I will then investigate automatic understanding and modelling of group-level social interactions (Tapus et al. 2019), including f-formations (Marshall et al. 2011), sociograms (as done in Garcìa-Magariño et al. (2016) for instance), and inter-personal affordances (Pandey and Alami 2013). This task builds on literature on social dynamics analysis, eg (Durantin et al. 2017; Jermann et al. 2009; Martinez-Maldonado et al. 2019), and apply it to real-time social assessment by a robot, itself embedded into the interaction.

Research strand 1.1 – targeted outcomes

A holistic model of humans, suitable for modelling human-robot and human-human interactions with high granularity; an algorithmic pipeline for the automatic analysis of social dynamics at group-level, able to model in real-time the social context of the robot.

1.2 – Social situation assessment

In 1.2, I integrate the social cues from 1.1 with traditional situation assessment platforms. It will result in a socio-cognitive model of the social environment of the robot that I term *social situation assessment*. It effectively extends the existing spatio-temporal representation capabilities of robots to the social sphere, and covers the development of a complete social assessment pipeline, from social signal perception (like automatic attention tracking, face recognition, sound localisation, etc.) to higher-level socio-cognitive constructs, including group dynamics, perspective taking (Flavell et al. 1992) and mutual modeling/mentalizing (Lemaignan and Dillenbourg 2015; Dillenbourg et al. 2016).

Research strand 1.2 – targeted outcomes

A novel cognitive sub-system for social situation assessment, integrated in an open-source software framework (see technical contributions page 18). This cognitive system will enable the robot to represent its physical and social environment, and perform queries about it, including queries about past events (temporal model) and queries requiring higher socio-cognitive perceptual capabilities like perspective taking and mentalizing.

1.3 – Social embeddings

One of the key novel scientific idea that I will research in this strand is the construction of a meaningful **social embedding** for the robot. *Embeddings*, as a machine learning concept, are compact, low-dimensional representations of much more complex phenomena. A well-known example of such embeddings is *word embeddings* (eg. *word2vec* (Mikolov et al. 2013) or *GloVe* (Pennington et al. 2014)) where words are projected into a n-dimensional space (with eg. $n = 300$ in the case of *GloVe*). Such an embedding can be constructed such as Euclidian distance in the embedding reflects semantic distance in the word-space, making it computationally inexpensive to represent and manipulate words' semantics. Similar embeddings are used in social signal processing, for instance for face (Schroff et al. 2015) or gesture (Ge et al. 2008) recognition.

I believe a similar approach could be devised for social interactions: constructing a low-dimensional embedding onto which social situations can be projected while they are perceived by the robot. Note that the additional complexity introduced by the dynamic nature of the interaction (ie, a social situation needs to be observed for some time to be correctly interpreted) could in principle be overcome using for instance a transformer architecture (Vaswani et al. 2017), as done for action recognition in eg. Girdhar et al. (2019).

If fruitful, this approach would significantly simplify the application of neural networks to automatically recognise social situations and social dynamics (something notoriously difficult to achieve with the current state-of-art, as I discuss in Bartlett et al. (2019)), and potentially *generate* plausible social situations, that the robot could use to eg. predict the next states of an interaction.

Research strand 1.3 – targeted outcomes

The investigation of *social embeddings* as a general, sub-symbolic representation of the social environment of interactive robots.

Focused experimental programme

In complement to the large-scale experimental work described in Strand 5, a focused experimental programme accompanies Strand 1, to specifically support and demonstrate the investigation of these socio-cognitive capabilities.

I intend implement a subset of the experimental protocols identified by Frith and Happé (1994) to investigate theory of mind with autistic children, as it offers an excellent experimental framework for social robotics, as I argued in Lemaignan and Dillenbourg (2015). Indeed, experimental protocols in research on autistic spectrum disorders are often striking by their apparent straightforwardness because of the careful choice of interaction modalities: since autistic children frequently exhibit impairments beyond social ones (such as motor or linguistic ones), the experiments must be designed such that they require only basic cognitive skills beyond the social abilities that are tested. This makes them especially well-suited for experimental robotics, properly isolating the cognitive competencies that we want to test and evidence early on in the development process of a complete cognitive architecture.

No mentalizing required	Mentalizing required
Ordering behavioural pictures	Ordering mentalistic pictures (Baron-Cohen et al. 1986)
Understanding see	Understanding know (Perner et al. 1989)
Protoimperative pointing	Protodeclarative pointing (Baron-Cohen 1989)
Sabotage	Deception (Sodian and Frith 1992)
False photographs	False beliefs (Leslie and Thaiss 1992)
Recognizing happiness and sadness	Recognizing surprise (Baron-Cohen et al. 1993)
Object occlusion	Information occlusion (Baron-Cohen 1992)
Literal expression	Metaphorical expression (Happé 1993)

Table 1: Tasks requiring or not mentalizing to pass, listed by Frith and Happé (1994)

Frith and Happé's list (Table 1) is in that regard especially interesting in that it mirrors pairs of task (ones which do not require mentalizing with similar ones which do require mentalizing), thus providing good control tasks. I intend to implement several of these tasks to support Strand 1's contributions to basic research in artificial social cognition – *Object occlusion* vs. *Information occlusion* (Baron-Cohen 1992), for instance, evidences *representation-level* perspective taking: adapting such a protocol for human-robot pairs would demonstrate *second-order*, *representation-level* perspective taking capabilities, which is beyond the state-of-the-art in an artificial cognitive system.

Strand 2: Generative social behaviours

Mirroring Strand 1's focus on understanding the social interactions, Strand 2 addresses the question of social behaviour *generation*: how to create natural behaviours, engaging over a sustained period of time (eg not simply picking scripted behaviours from a library, that are rapidly perceived as repetitive).

The focus of my research will be in the first instance on *non-verbal* behaviours. This is a purposeful interaction design choice, that ensures we can more effectively manage what cognitive capabilities are ascribed to the robot by the users (expectation management). I seek however to significantly push forward the state-of-the-art of behaviour generation for robots, both in term of technique to generate the behaviours, and in term of the nature of the non-verbal behaviours (including expressive gestures and motion, non-verbal utterances using sounds, gaze, joint attention).

2.1 – Design of novel interaction modalities

As part of Strand 2, I intend to lead research on novel non-verbal interaction modality for social robots. I will pursue an interdisciplinary, bottom-up approach involving artists and field practitioners, exploring new communication modalities like creating a body language for social robot with choreographers, or investigating new forms of sound expressions like soundscapes with sound experts (soundscapes are sound environments that reflects a particular situation; they have been shown to be an effective intervention technique in the context special needs treatments, eg Greher et al. (2010)).

I specifically want to question the traditionally uncreative approach of HRI where human codes of communication are typically replicated 'as it' on robots. To discover new modes of interaction, potentially better-suited to sustain long-term engagement, I want to invite artists and creators of diverse backgrounds to create new, original communication codes (like I did with comedians in the *roboscopie* theatre play (Lemaignan et al. 2012)).

As an illustration, Figure 3 shows an immersive setup where a dancer or a choreographer 'takes control' of the robot's body via motion capture, while seeing through the robots' eyes (VR headset or a CAVE projection, similar to (Bailly et al. 2015)). For a period of time, the artist 'wizard' is tasked to invent a body language for the robot while interacting with the humans it meets, by remotely controlling the robot. This would result in a large dataset of freely created, socially-congruent physical behaviours, that would directly feed into eg Generative Adversarial Networks, as presented below.

Research strand 2.1 – targeted outcomes

The research, development and implementation of novel non-verbal communication modalities, including for instance a robot 'body language' for social interactions and soundscapes; a large dataset of such interactions, recorded in immersive conditions, and suitable for machine learning.

2.2 – Generative neural network for social behaviour production

Designing behaviours that enable sustained, long-term engagement in a social human-robot interaction is essentially an open research question. The specific challenge of producing non-repetitive social behaviours is particularly difficult: social robots typically rely on *off-the-shelf behaviours*, where the robot effectively picks from a set library of behaviours (that might be individually relatively complex). The approach can elicit a strong initial social response from the user, but this social response tends to vanish rapidly once the 'tricks' of the robot have been all discovered and become repetitive. Besides, as the robot does not typically maintain a long-term socio-cognitive plan of the interaction, the behaviours are typically perceived as fun, yet pointless, leading to disengagement. This is often observed in toy-like robots (eg Vector, Dot & Dash) (Hoffman 2019).

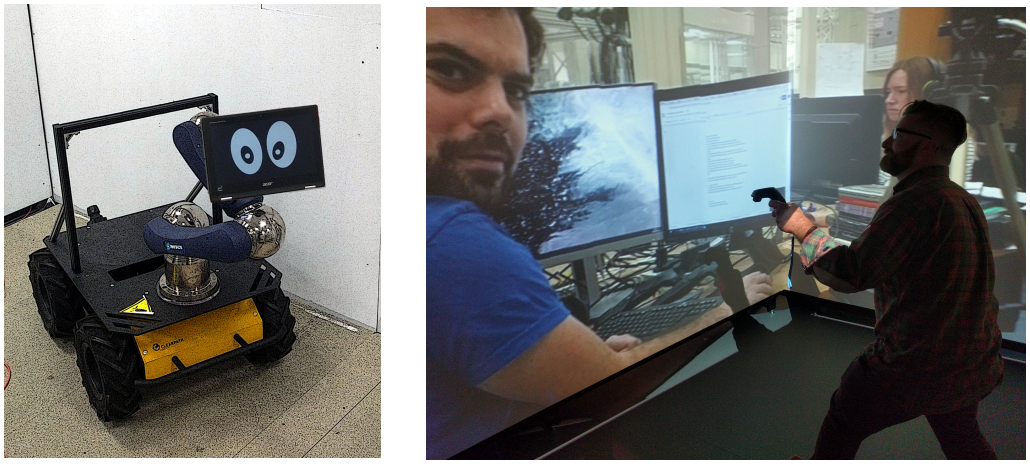

Figure 3: (left) Possible appearance of a puppet-robot that I will use to collect data. A tablet, displaying facial animations, is mounted on a robotic arm. It can freely orient its 'gaze' and use expressive movements. The robot is effectively teleoperated in realtime by an artist (eg choreographer, right) who 'sees through the robots' eyes' and who is tasked with designing and acting a novel 'body language' for social interactions.

While effectively enabling the robot to store and manage long-term plans, symbolic task planners still rely on mostly static libraries of 'canned', repetitive actions. Also neither of these planners are well-suited to rapid, dynamic behaviours generation, especially in situations requiring performing parallel, blended actions.

Building on these solid foundations, I aim at significantly advancing the state of the art in this regard, by combining two recent machine-learning techniques: (1) generative neural networks for affective robot motion generation (Yang and Peters 2019; Marmpena et al. 2019; Suguitan and Hoffman 2020); (2) interactive machine learning in high-dimensional input/output spaces, where I have shown with my students promising results for generating complex social behaviours (Senft et al. 2019; Winkle et al. 2020b) that fully involve the end-users (Winkle et al. 2018).

In Suguitan and Hoffman (2020), a Generative Adversarial Network (GAN) is trained to generate expressive motions; the generation being modulated by a feature encoding an emotion. I will extend this idea in two ways: (1) I will train the GAN on multiple interaction modalities (motions, but also facial expressions, gaze, sounds) using the data acquired in 2.1. The aim will be to collect a large amount of data to train a GAN from, effectively creating a new multi-modal 'grammar' for the robot expression. (2) Instead of using emotions to modulate the generation stage, I will use the social embedding constructed in 1.5: the generated behaviours will be shaped by the current, complex social state of the interaction instead of simply emotions.

Research strand 2.2 – targeted outcomes

A generative neural network able to produce non-verbal yet multi-modal social behaviours. They will combine expressive gestures, gazing behaviours, facial expressions, and expressive sounds.

Strand 3: Goal-driven socio-cognitive architecture

Strand 3 investigates the principled integration of a cognitive architecture for autonomous social robots. It binds together the socio-cognitive perceptual capabilities of the robot developed in the Strand 1, the action production mechanisms developed in the Strand 2, and includes key elements

from my 'human-in-the-loop' methodology to isolate and model the interaction principles and social goals of the robot.

3.1 – A social teleology for robots

Teleological systems (ie goal-driven) have been investigated in robotics for being a way of providing long-term drives to an autonomous robot. This has been successfully applied to relatively simple cognitive systems (Oudeyer et al. 2005; Moulin-Frier et al. 2014) or virtual agents (Pathak et al. 2017). This first basic research activity in Strand 3 aims to significantly progress this line of research, and to look into *complex* interactive cognitive systems. The key objective of this work package 3.1 is to define and implement a novel *social teleology*: the **algorithmic encoding of long-term social goals** into the robot.

This work will directly draw from the participatory, 'human-in-the-loop' methodological paradigms that present in Strand 5. Indeed, before being transposed into algorithms, these long-term social goals will first be co-defined and co-created by the end-users and the public in terms of *interaction principles for useful and responsible social robots*.

Research strand 3.1 – targeted outcomes

The algorithmic translation of interaction principles into long-term social goals for the robot; eg a long-term, socially-driven action policy for the robot.

3.2 – Learning from humans to achieve by-design responsible & trustworthy AI

I have recently obtained promising results on human-in-the-loop social learning (Senft et al. 2019; Winkle et al. 2020b): non-expert end-users teach in-situ (eg at school, at the gym, etc.) a robot, which progressively learns to be autonomous, eventually reaching full task- and social autonomy. This approach, that I developed with one of my students (Senft et al. 2017), holds a lot of promise in term of field acceptance of social robots as it entrust the end-user with a high level of control during the learning phase, leading to a feeling of ownership of the resulting robot behaviours. I will further develop this idea, applying in-situ interactive reinforcement learning to more complex, real-world, situations.

In addition, I will study through qualitative methods (thematic interviews and questionnaires) whether (and how) human-in-the-loop machine learning enables a more trustworthy AI system, by involving the end-users in the creation of the robot behaviours, thus offering a level of behavioural transparency to the end-users.

Research strand 3.2 – targeted outcomes

A human-in-the-loop reinforcement learning paradigm, suitable for in-situ teaching of the robot by the end-users themselves, demonstrated in complex social environments.

3.3 – Integrating a socially-driven architecture for long-term interaction

This research strand builds on the state of art in cognitive architectures (disembodied ones (Chong et al. 2007; Vernon et al. 2007; Kingdon 2008; Duch et al. 2008; Langley et al. 2009; Taatgen and Anderson 2010; Thórisson and Helgasson 2012), as well as ones specifically developed for robotics: ACT-R/E (Trafton et al. 2013), HAMMER (Demiris and Khadhouri 2006), PEIS Ecology (Saffiotti and Broxvall 2005; Daoutis et al. 2012), CRAM/KnowRob (Beetz et al. 2010; Tenorth and Beetz 2009), KeJia (Chen et al. 2010), POETICON++ (Antunes et al. 2016), and the LAAS Architecture for Social Interaction (Lemaignan et al. 2017), to which I have been a key contributor during my PhD). The

overall purpose of this socio-cognitive architecture is to integrate in a principled way the spatio-temporal and social knowledge of the robot (Strand 1) with a decision-making mechanism, to eventually produce socially-suitable actions (Strand 3).

The decision-making mechanism is critical, and lay at the heart of my research project. The robot will rely on it to generate action decision that are purposeful, legible and engaging on the long run, something that none of the existing architectures have been able to successfully demonstrate to date. I aim at a breakthrough, and will introduce a novel approach: drawing from the interaction patterns identified (Strand 4), I will combine long-term, socially-driven goals (the *social teleology*, 3.1), and human-in-the-loop machine learning (3.2) using a novel arbitration mechanism.

The arbitration mechanism itself will build on research on reinforcement learning for experience transfer (Madden and Howley 2004) that enables the re-assessement of a policy (here, our long-term social teleology) based on specific experience (here, the end-user-taught policy).

Research strand 3.3 – targeted outcomes

A socio-cognitive architecture, fully implemented on interactive robotic platfoms, that enables long-term social engagement, by combining long-term goals with domain-specific action policies, taught by the end-users themselves.

Strand 4: Framing responsible 'robot-supported human-human interactions'

The basic, long-term ambition of my research programme is to re-investigate the underpinnings of human-robot interaction by taking a **strong human-centered perspective**. I frame this as a shift from *human-robot interaction* to *robot-supported human-human interactions* (r-HHI). This last major strand of research operationalises this objective in term of a basic contribution: examining the interplay between r-HHI, responsible AI, and ethics. This will be directly influenced by the experiemental work described in Strand 5.

Conceptual framing of r-HHI and ethical framework

The first task in Strand 4 is to research and define the framework that will provide the conceptual frame around questions like: what role should social robots have? Where to set the boundaries of artificial social interactions? What does 'ethical-by-design', 'responsible-by-design' mean in the context of social human-robot interactions?

Indeed, my research project involves social robots, interacting in repeated ways and over long period of time, with human end-users, including vulnerable children. This raises complex ethical issues, both practical ones (how to design the experimental work in a such a way that they are safe and ethically sound), and more fundamental ones (what is the ethical framework for robots intervening in socially sensitive environment?).

The ethical questions raised by social robotics have been actively studied over the last 5 years, attempting to address issues like:

- how to ensure that social robots are not used to simply replace the human workforce to cut costs?
- can we provide guarantees that the use of social robots will always be ethically motivated?
- further on, can we implement some ethical safeguarding built-in the system (like an ethical *black-box* (Winfield and Jirotka 2017))?
- what about privacy? how to trust robots in our home or school or hospital not to eavesdrop on our private lives, and, in the worst case, not be used *against* us?

These questions are indeed pressing. The recent rise of personal assistants like Amazon Alexa or Google Home, with the major privacy concerns that accompanies their deployments in people home, shows that letting the industry set the agenda on these questions is not entirely wise – and robots can potentially be much more intrusive than non-mobile smart speakers. The EU is positioning itself at the forefront of those questions. The recent release of operational **Ethics Guidelines for Trustworthy AI** by the EU High-level Expert Group on Artificial Intelligence (*Ethics Guidelines for Trustworthy AI* 2019) is a strong sign of this commitment. These guidelines identify seven requirements of trustworthy AI:

- **R1** **Human agency and oversight**, including fundamental rights, human agency and human oversight
- **R2** **Technical robustness and safety**, including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- **R3** **Privacy and data governance**, including respect for privacy, quality and integrity of data, and access to data
- **R4** **Transparency**, including traceability, explainability and communication
- **R5** **Diversity, non-discrimination and fairness**, including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- **R6** **Societal and environmental wellbeing**, including sustainability and environmental friendliness, social impact, society and democracy
- **R7** **Accountability**, including auditability, minimisation and reporting of negative impact, trade-offs and redress.

The design methodologies and techniques employed in my research programme naturally implement most of these requirements: interaction co-design and human-in-the-loop machine learning ensures human agency oversight over the robot's behaviours (R1); Privacy and data governance (R3) is addressed in the project's data management plan and facilitated by the design decision of performing all data processing on-board the robot, avoiding the dissemination of personal information; the transparency of the robot behaviour (R4) stems from the machine learning approach that we advocate: the robot's behaviours primarily originate from what the end-users themselves taught the robot; diversity and non-discrimination (R5) is supported by the large-scale involvement of the public at the science centre, ensuring a broad diversity of backgrounds and profiles; societal wellbeing (R6) is the core research question of the project, and I intend to contribute in realising this requirement in the context of social robots.

Technical robustness (R2) and accountability (R7) are important design guidelines for the robot's cognitive architecture (WP4), and will be addressed there as well.

The Ethics Guidelines for Trustworthy AI form a solid foundation for the project. However, personal and social robots raise additional questions regarding what ethical and trustworthy systems might look like, and while the principles of responsible design are somewhat established (Stahl and Coeckelbergh 2016; BSI 2016), the reality of robot-influenced social interactions is not fully understood yet, if only because the technology required to experience such interactions is only slowly maturing.

Social robots have indeed two properties that stand out, and distinguish them from smart speakers, for instance. First, they are fully embodied, and they physically interact with their environment, from moving around, to picking up objects, to looking at you; second, willingly or not, they are ascribed *agency* by people. This second difference has far-reaching consequences, from affective bonding to over-trust, to over-disclosure of personal, possibly sensitive, informations (Martelaro et al. 2016; Shiomi et al. 2017). As an example, a common objection to human-

robot interaction is the perceived deceptive nature of the robot's role. It has been argued (Bisconti Lucidi and Nardi 2018) that the underlying concern is likely the lack of an adequate (and novel) model of human-robot interactions to refer to, to which the project will provide elements of response. This needs nevertheless to be accounted for in depth.

Ethical framing of social robotics has started to emerge under the term **roboethics**: the "subfield of applied ethics studying both the positive and negative implications of robotics for individuals and society, with a view to inspire the moral design, development and use of so-called intelligent/autonomous robots, and help prevent their misuse against humankind." (Allen et al. 2011). Specific subfields, like assistive robotics (Sharkey and Sharkey 2012), have seen some additional work, but social robotics is still not equipped with operational guidelines, similar to the EU guidelines on trustworthy AI.

This is my ambition that my research will significantly contribute to the framing and the building of guidelines and recommendations for responsible human-robot interactions, enabling a safe and trustworthy digital future in our society.

I intend to develop this line of research by adopting the same open-science approach, involving the general public in the process: my field work experiments (Strand 5) will both *build on* and *feed into* the framework developed in this research strand.

In addition, I will also structure this framing effort through international workshop. During these workshops, invited ethics experts (from both robotics and AI backgrounds) will be invited to engage with field practitioners, including local institutions where the experimental work will have taken place. These regular forums will be coordinated with the European Commission (I am already acting as an Expert Collaborator on ethics for the European Joint Research Centre), and will create a public opportunity to debate and iterate over ethics guidelines for responsible long-term social interactions with robots.

Research strand 4 – targeted outcomes

A conceptual framework that clarify and organise together the questions raised by long-term social interactions; ethical guidelines for such interactions, aimed at informing future policy making.

Strand 5: Experimental programme: long-term deployments in complex social spaces

I intend to scaffold and demonstrate the results of my research with an ambitious experimental programme. This experimental programme is two-fold: (1) informing the design of desirable social robots, by creating a 'crowd-scale' methodology to co-design and co-construct the robots' social role and behaviours; (2) demonstrating meaningful progress on robot-supported social interventions, in complex and high-impact real-world environments.

5.1 – Crowd-sourced patterns of robot-supported social interactions

In order to broadly engage the public with defining what future robots should do to be perceived as responsible, beneficial, and engaging, I intend to create and deploy a novel investigation methodology that I term 'experimental crowd-sourcing'. For a whole year, and in close partnership with local institutions (like Toulouse's "Le Quai des Savoirs"), I will deploy the laboratory's robot(s) within a public space, relinquishing its control to the visitors themselves. Tasked with remotely operating the robot to assist fellow visitors, a researcher will accompany them in 'inventing by doing' a new grammar of social interactions to develop answers to the questions: what does it

mean for a robot to help? How to do so in the dynamic, messy, public environments? What are acceptable behaviours? Can we see new social norms emerge? At the end of this experiment, we expect 1000s of people to have experienced – and co-designed – how robots should interact with humans in a positive, helpful way. Each of these experiences will contribute to uncovering and designing the basic principles of social interaction for robots.

This work will be based on my previous research on 'human in the loop' machine learning (Senft et al. 2019; Winkle et al. 2020b), while scaling it up both in term of duration (several months) and volume of input (hundreds of participants).

Research strand 5.1 – targeted outcomes

A set of crowd-sourced interaction patterns and principles, that will inform the long-term social goals of the robot (4.1); a large dataset of social interactions to feed into Strands 2 and 3.

5.2 – Experimental fieldwork

As mentioned in my track-record, I have 7+ years of experience deploying robots 'in the wild' (including schools, gyms, medical surgeries) where robots are used by non-experts in ecologically valid environments. When conducted with the highest standards of scientific rigour, these deployments provide invaluable insights on the actual challenges that prevent broader acceptance and use of robots in the society.

In the first 5 years of starting my role at CNRS, I would strive to demonstrate my approach with two such ambitious field experiments.

The first one would involve the deployment of social robots in local special needs schools (SEN schools), building on the initial pilots that I have conducted at the Bristol Robotics Lab. Building on a rigorous participatory approach involving the school teachers, as well as the parents, we would seek to integrate robots in the daily life of the schools, supporting the development of the students' physical and social skills.

The second one could take place at Toulouse Children's Hospital, supporting isolated children who suffer long-term conditions, in close cooperation with the hospital staff. Likewise, social robots would be deployed on premises, for one uninterrupted year. They would integrate the daily routines of the institution, under supervised autonomy (Senft et al. 2017), and *without* requiring or expecting the presence of a researcher at all time.

These two examples raise specific practical and ethical questions, as they target vulnerable populations. This is however an informed choice: with their complex social situations and social dynamics, they would allow to convincingly demonstrate the importance and positive impact of socially-driven, socially-responsible robotics. When implemented, these scenarios would also actually deliver high societal impact, with tenths of children to directly benefit of the project. This would show how robots can have a lasting, beneficial impact on the society, alongside human carers, and it will establish the idea of *robots supporting human interactions* instead of dehumanising our social relationships.

Importantly, these deployments would take place within the strict ethical framework established in Strand 4, providing real-world feedback on the suitability and thoroughness of the ethical guidelines.

Developing methodological and technical excellence

Finally, I would like to sketch out my intended contributions to one key underpinning of my research programme: technical and methodological excellence. Indeed, it is important to re-state

here the fundamental nature of my project: **the development of novel, fully autonomous, social robots, able to foster long-term interactions**. This aim is beyond the state of the art, and while ambitious science is required to achieve it, **technical excellence is the fundamental enabler**. I am excited about such a major technical challenge, and following on from my strong technical track-record, I will innovate to build and assemble the leading technology required to bring autonomous social robots to the general public.

Advancing open-source software for HRI

As outlined in the presentation of my past contributions, I have a long academic track-record of technical open-source contributions to robotic algorithms and software; to name a few significant ones: the `oro` knowledge base (Lemaignan et al. 2010), the natural language processing with semantic grounding tool `dialogs` (Lemaignan et al. 2011), the 3D situation assessment platform `underworlds` (Lemaignan et al. 2018c), the LAAS architecture for social interaction (Lemaignan et al. 2017), or new algorithms for interactive reinforcement learning (Senft et al. 2017)). I intend to pursue these efforts, also embracing and integrating novels techniques from neighbouring fields. Indeed, astonishing progress has been made over the past ten years in AI, largely due to the success of machine learning techniques to classify complex signals. As highlighted in the previous sections, I intend to fully embrace these techniques when and where relevant (for instance, *social embeddings* to recognise complex social situations or generative networks to create on the fly complex social gestures).

One critical aspect of the technology that I want to investigate in depth is **real-world robustness**: indeed, many of the impressive achievements of machine learning do not hold well when applied to messy, real-world situations. While *not* an intrinsic weakness of machine learning, it reflects a lack of focus and experience with the more challenging experimental conditions encountered when robots are deployed outside of the labs. As presented in Strand 5, I have an ambitious experimental programme, which will provide complex real-world data to train next-generation deep neural networks, suitable for robust social perception and action generation.

I will carry on this commitment to developing new software and algorithms with a special attention to the software engineering challenges that are often overlooked by academics and, as a result, hamper dissemination: for instance, continuous integration, packaging, automated testing using simulation, deep integration with existing frameworks and standards (especially ROS).

The challenges raised by (the lack of) real-world robustness and software complexity are especially visible when building social perception pipelines (the main goal of my research strand 1). It effectively requires to combine together tenths of specialist software components into one coherent and principled framework, with algorithmic redundancy, and complex testing requirements. I have initiated this work (Figure 4), with preliminary steps in place for multi-modal human modeling. However, providing real-world robustness and overcoming training sets weaknesses and biases will be a long-term effort that I intend to spearhead as a CNRS researcher.

Future-proofing robotic platforms for interaction

My research programme focuses on the AI engine of social robot, rather than eg hardware development. As such, my work relies on pre-existing platforms, suitable for social human-robot interactions.

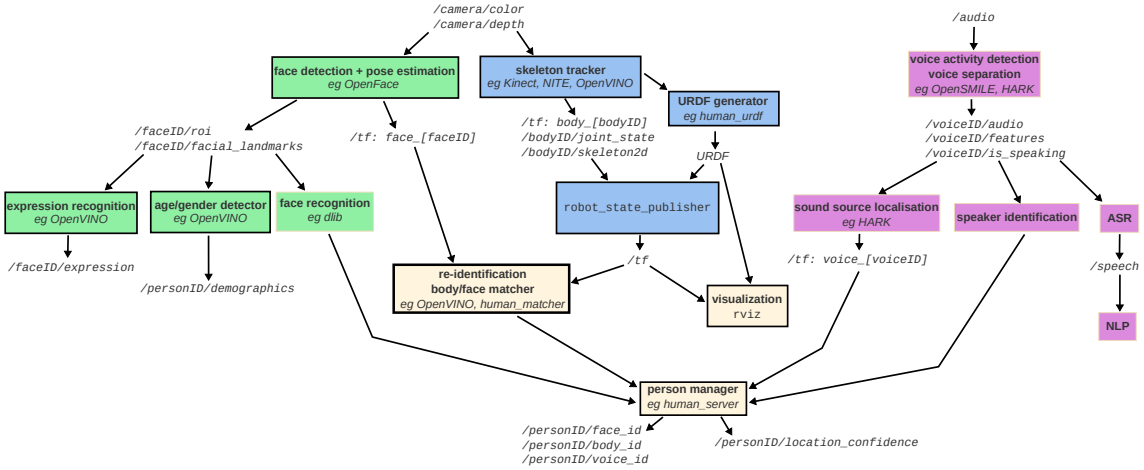

Figure 4: Overview of the 'ROS4HRI' framework (Mohamed and Lemaignan 2020) that I designed and whose development I currently supervise.

However, looking specifically at human-sized mobile manipulators with advanced social features, the choice of robotic platforms is in effect limited. The two leading mobile social robots available on the market today, SoftBank Pepper and PAL TiaGo (along with the Fetch Mobile Manipulator, which is functionally similar to PAL's TiaGo) are 5+ years old design. While they have played an important role in supporting research in HRI over the past years, they do not reflect the current state of the art in social mobile manipulators , and as a CNRS researcher, I will seek internal and/or external funding to acquire state-of-the-art platforms like the HALODI Eve or the IIT R1 (Figure 5), eg via joint applications to EU funding.

Fostering open-science best practices

Finally, I see fostering **open science best practices** as a cornerstone of technical and methodological excellence. I am a long-standing open science proponent and I have led a number of targeted actions to further develop these practices: the publication and an analysis of the current practices and recommendations in field of human-robot interaction (Baxter et al. 2016); the publication of large open datasets (Kennedy et al. 2016; Lemaignan et al. 2018a; Lemaignan et al. 2018b); open access publishing; a large corpus of open-source contributions (180+ public repositories on Github), including key contributions to major research software like ROS (I led the port of ROS to Python3); active and open engagement with the wider community via social media.

More can be done, however, and as a CNRS researcher, I would commit to further enhance the

I will leverage my editorial role in high-profile publications and conferences (most notably FrontiersIn Robotics and AI, and the ACM/IEEE Human-Robot Interaction conference) to advance this agenda.

- **reproducible science** by always providing the code and data analysis scripts used in experiments;

- **pre-registered studies** where the exact methodology is published *before* performing the study itself, ensuring the highest standard of experimental rigour;
- **continued lobbying** in international academic venues to support and foster open-science best practices, in particular by opening the Human-Robot Interaction conference to pre-registered studies

finish section

Proposed host laboratories

Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS) – UPR 8001

The Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS), part of the *Artificial and Natural Intelligence Toulouse Institute* (ANITI), has a long and strong track-record in autonomous interactive robots. Combined with the breadth of expertise in AI available within the Toulouse-wide ANITI institute, it would effectively support and accelerate several of the key science breakthrough I target with this project.

Indeed, the LAAS is one of the few research group worldwide that has achieved full autonomy for complex socio-cognitive robots (eg Lemaignan et al. 2017). This has been made possible by the long-term commitment of the laboratory to cognitive robotics, spearheaded by Pr. Rachid Alami. Over the last 20 years, Alami and colleagues have indeed built a large, ambitious body of research, enabling a wide range of cognitive skills on service robots. To cite a few that would be directly relevant to my research programme: reasoning languages (Ingrand et al. 1996) and supervision (Clodic et al. 2009); 3D motion and task planning (Sisbot et al. 2008; Mainprice et al. 2011); human-aware symbolic task planning (Alami et al. 2005; Alili et al. 2008; Lallement et al. 2014; Milliez et al. 2016); cognitive architectures (Devin and Alami 2016).

The LAAS also has a long-standing commitment to technical excellence. This has enabled a large number of technically challenging experimental deployments of robots, including social and interactive robots (eg (Alami et al. 2005)). This expertise would directly benefit my research programme as well.

My research programme links directly to these themes, and focuses additionally on (1) social AI, (2) data-driven human-robot interaction, and (3) the real-world societal impact of human-robot interactions. Those emerging research areas are not yet developed at LAAS; my affectation there would bring this novel expertise.

At the local level, my interdisciplinary research programme would fits especially well in the ANITI agenda (one of the four French Institute for AI; 200+ researchers in the Toulouse region): as put by Nicholas Asher, ANITI's director, one of the ANITI's challenges is to identify powerful transverse applications, that would enable the integration of the range of AI techniques developed within the institute into a complete system. The robots I will develop in my programme would be natural candidates for such integrations, where research on eg Work on Fair & Robust Machine Learning (Jean-Michel Loubès), Social Trust in AI (Céline Castets-Renard), AI for policy learning (Cesar Hidalgo, Leïla Amgoud), or AI for complex behaviour generation (Nicolas Mansard), could be combined.

Institut des Systèmes Intelligents et de Robotique (ISIR) – UMR 7222

TODO Section

Importance, interdisciplinarity and conclusion

National and International Importance

This research project addresses the questions of how to design socially assistive robots that are both effective autonomous social agent, and useful, acceptable and responsible vis-à-vis their end-users.

From an academic perspective, France and the European Union currently enjoy a 2-3 years leadership on research and deployment of socially interactive robots, mainly built through the several large-scale European projects on that topic, which took place over the last decade. The CNRS did play a key role in several of these projects, eg FP6-Cogniron, FP7-CHRIS, H2020-Spencer, H2020-MuMMer, and has built a solid reputation. It is now critical that this expertise is maintained and further developed, as to ensure continued future academic leadership of the CNRS, France and the European Union in this fast developing, socially critical, domain.

Indeed, surprisingly few groups worldwide have achieved full autonomy for a complex social robot, the LAAS-CNRS being such a rare examples. **By joining the CNRS, I will create the conditions to 'future-proof' this scientific know-how, both consolidating our expertise, and leading a wide-ranging set of new research directions on social robotics**. I will as such contribute to further **assert scientific leadership on these socially-transformative technologies**.

In addition, my project would create the opportunity for France and Europe to establish themselves at the forefront of the emerging research on the complex ethical questions arising from the development of social robots. Indeed, my research will significantly contribute to the pressing issues around Responsible AI applied to robotics: the creation of the High-level Expert Group on Artificial Intelligence by the European Union, and the subsequent release in 2019 of their *Ethics guidelines for trustworthy AI*, evidences the importance of framing and defining the adequate policies to enable and support the future development of a safe and trustworthy AI. It however does not address any of the emerging challenges raised by social robots.

My work will in effect pave the way for similar guidelines to be extended to social robotics, eg, *embodied, physical* AI. In line with the Europe Union's strong societal values, the project specifically addresses and framees the ethical underpinnings of social robots and deliver the guidelines that we need to inform our future policies on social robotics. Combined with beyond-state-of-the-art technological developments, **this research programme will make a major contribution in securing a safe and responsible digital future in France, the European Union, and beyond**.

Interdisciplinary nature of the research programme

This research programme paves the way for a better understanding of the societal challenges raised by the rapid development of AI and robotics. Grounded in both the psycho-social literature of human cognition, and the latest technological advances in artificial cognition and human-robot interaction, the project delivers major conceptual, technical and experimental contributions across several fields: AI, ethics, sociology of technology, socio-cognitive psychology, intelligent robotics, learning technology. As such, **my research project builds bridges across multiple disciplinary boundaries**.

I intend to deliver this programme by building on a range of multidisciplinary methods, including user-centered design; ethnographic and sociological investigation; expressive non-verbal communication taking inspiration from the arts (dance and puppetering); embodied cognition; symbolic AI; neural nets and sub-symbolic AI; interactive machine learning.

Accordingly, I intend to significantly extend the **strong interdisciplinary links** that have al-

ready been established in the Human-Robot Interaction groups at both LAAS-CNRS and ISIR. I will seek funding to recruit PhDs and post-docs with diverse backgrounds in sociology of technology, cognitive modeling, machine learning, cognitive robotics. Additional expertise will be sought through academic collaborations: the creation of an Ethics of HRI working group will contribute expertise to guide the work on ethics; specific collaboration with researchers and practitioners in education and psychology will provide expertise in learning technologies and cognitive impairment; collaborations with artists (dancers, sound artists) will provide additional expertise and insights on expressive communication; and collaboration with local institutions (eg science charities, schools, hospitals) will complete the **open and interdisciplinary culture that I aim to foster** within the host laboratory.

Conclusion: my scientific vision

My research programme is ambitious, both in the short term, and in the longer run: **I will lead the design, implementation and real-world demonstration of socially-intelligent robots. My aim is to create, sustain and better understand the dynamics of responsible long-term social human-robot interactions, in order to build robots that (1) have an effective, demonstrable social utility, and (2) will see long-term acceptance by their end-users**.

In the short term (next 5 years), I will bring together two emerging AI paradigms (teleological architectures and human-in-the-loop machine learning); I will integrate them into a state-of-the-art cognitive architecture for autonomous social robots, relying on multidisciplinary approaches where relevant; I will create the conditions for a unique, large-scale, 'public-in-the-loop' participatory design approach that will transform how we think about public engagement with robotic design; finally, I will co-design and deploy an autonomous robot in a real-world, highly social setting, demonstrating social usefulness and acceptance over a significant period of time.

In the longer run, I will drive, locally, nationally and internationally, the future of intelligent social robots. Both from a societal point of view (what role for robots in our society? how to involve the general public into the design and implementation of these robots? how to ensure the technology is inclusive? what ethical framework?) and from a technological point of view (what models of the humans and their environment do we need to build? what algorithms for robots to 'learn by doing' and become 'good citizens' in our digital society?)

These two facets (societal progress and technical progress) should always go hand-in-hand: societal research needs to fully understand the limits and opportunity of the underlying technology; technology must be framed by societal needs and ethical considerations. I believe my cross-disciplinary academic profile and extensive experimental and technological experience provide me with the skills and understanding required to successfully progress them both.

At its core, my research programme is about co-defining, co-designing and building the social robots of tomorrow: it offers a vision of AI and social robotics that places the human at the centre of these emerging technologies, to foster novel social dynamics that are acceptable and beneficial to society.

I propose an ambitious yet realistic scientific and technical pathway to progress toward this major scientific endeavour. By joining the CNRS, I will lead the creation of autonomous social robots that not only learn social behaviours with and from the public and end-users, but that are also co-designed from the ground-up to be acceptable, responsible and useful to the humans they will serve.

References

[1] R. Alami, A. Clodic, V. Montreuil, E. A. Sisbot, and R. Chatila. “Task planning for human-robot interaction”. In: *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*. 2005, pp. 81–85.

[2] S. Alili, V. Montreuil, and R. Alami. “HATP Task Planer for Social Behavior Control in Autonomous Robotic Systems for HRI”. In: *The 9th International Symposium on Distributed Autonomous Robotic Systems*. 2008.

[3] C. Allen, W. Wallach, J. J. Hughes, S. Bringsjord, J. Taylor, N. Sharkey, M. Guarini, P. Bello, G.-J. Lokhorst, J. van den Hoven, et al. *Robot ethics: the ethical and social implications of robotics*. MIT press, 2011.

[4] A. Antunes, L. Jamone, G. Saponaro, A. Bernardino, and R. Ventura. “From Human Instructions to Robot Actions: Formulation of Goals, Affordances and Probabilistic Planning”. In: *ICRA*. 2016.

[5] L. Baillie, C. Breazeal, P. Denman, M. E. Foster, K. Fischer, and J. R. Cauchard. “The challenges of working on social robots that collaborate with people”. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–7.

[6] G. Bailly, F. Elisei, and M. Sauze. “Beaming the gaze of a humanoid robot”. In: *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. 2015, pp. 47–48.

[7] S. Baron-Cohen. “Out of sight or out of mind? Another look at deception in autism”. In: *Journal of Child Psychology and Psychiatry* 33.7 (1992), pp. 1141–1155.

[8] S. Baron-Cohen. “Perceptual role taking and protodeclarative pointing in autism”. In: *British Journal of Developmental Psychology* 7.2 (1989), pp. 113–127.

[9] S. Baron-Cohen, A. Leslie, and U. Frith. “Mechanical, behavioural and intentional understanding of picture stories in autistic children”. In: *British Journal of developmental psychology* 4.2 (1986), pp. 113–125.

[10] S. Baron-Cohen, A. Spitz, and P. Cross. “Do children with autism recognise surprise? A research note”. In: *Cognition & Emotion* 7.6 (1993), pp. 507–516.

[11] M. Bartlett, C. E. R. Edmunds, T. Belpaeme, S. Thill, and S. Lemaignan. “What Can You See? Identifying Cues on Internal States from the Kinematics of Natural Social Interactions”. In: *Frontiers in AI and Robotics* (2019). DOI: 10.3389/frobt.2019.00049.

[12] P. Baxter, J. Kennedy, S. E., S. Lemaignan, and T. Belpaeme. “From Characterising Three Years of HRI to Methodology and Reporting Recommendations”. In: *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference (alt.HRI)*. 2016. ISBN: 978-1-4673-8370-7. DOI: 10.1109/HRI.2016.7451777.

[13] M. Beetz, L. Mösenlechner, and M. Tenorth. “CRAM — A Cognitive Robot Abstract Machine for Everyday Manipulation in Human Environments”. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2010.

[14] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka. “Social robots for education: A review”. In: *Science robotics* 3.21 (2018), eaat5954.

[15] P. Bisconti Lucidi and D. Nardi. “Companion Robots: The Hallucinatory Danger of Human-Robot Interactions”. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’18. New Orleans, LA, USA: ACM, 2018, pp. 17–22. ISBN: 978-1-4503-6012-8. DOI: 10.1145/3278721.3278741. URL: http://doi.acm.org/10.1145/3278721.3278741.

[16] M. Bruckner, M. LaFleur, and I. Pitterle. “Frontier issues: The impact of the technological revolution on labour markets and income distribution”. In: *Department of Economic & Social Affairs, UN* 24 (2017).

[17] BSI. *Robots and robotic devices: Guide to the ethical design and application of robots and robotic systems*. Tech. rep. BS 8611:2016. BSI Standards Publication, 2016.

[18] X. Chen, J. Ji, J. Jiang, G. Jin, F. Wang, and J. Xie. “Developing high-level cognitive functions for service robots”. In: *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*. AAMAS ’10. Toronto, Canada: International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 989–996. ISBN: 978-0-9826571-1-9.

[19] H.-Q. Chong, A.-H. Tan, and G.-W. Ng. “Integrated cognitive architectures: a survey”. In: *Artificial Intelligence Review* 28.2 (2007), pp. 103–130.

[20] C. Choy, J. Gwak, and S. Savarese. “4D spatio-temporal convNets: Minkowski convolutional neural networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3075–3084. DOI: 10.1109/CVPR.2019.00319.

[21] A. Clodic, H. Cao, S. Alili, V. Montreuil, R. Alami, and R. Chatila. “SHARY: A Supervision System Adapted to Human-Robot Interaction”. In: *Experimental Robotics: The Eleventh International Symposium*. Ed. by O. Khatib, V. Kumar, and G. J. Pappas. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 229–238. ISBN: 978-3-642-00196-3. DOI: 10.1007/978-3-642-00196-3_27.

[22] A. Coninx, P. Baxter, E. Oleari, S. Bellini, B. Bierman, O. B. Henkemans, L. Cañamero, P. Cosi, V. Enescu, R. R. Espinoza, et al. “Towards long-term social child-robot interaction: using multi-activity switching to engage young users”. In: *Journal of Human-Robot Interaction* 5.1 (2016), pp. 32–67.

[23] M. Daoutis, S. Coradeschi, and A. Loutfi. “Cooperative knowledge based perceptual anchoring”. In: *International Journal on Artificial Intelligence Tools* 21.03 (2012), p. 1250012.

[24] Y. Demiris and B. Khadhouri. “Hierarchical attentive multiple models for execution and recognition of actions”. In: *Robotics and autonomous systems* 54.5 (2006), pp. 361–369.

[25] D. Dereshev, D. Kirk, K. Matsumura, and T. Maeda. “Long-Term Value of Social Robots through the Eyes of Expert Users”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019. ISBN: 9781450359702. DOI: 10.1145/3290605.3300896. URL: https://doi.org/10.1145/3290605.3300896.

[26] S. Devin and R. Alami. “An implemented theory of mind to improve human-robot shared plans execution”. In: *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press. 2016, pp. 319–326.

[27] P. Dillenbourg, S. Lemaignan, M. Sangin, N. Nova, and G. Molinari. “The Symmetry of Partner Modelling”. In: *Intl. J. of Computer-Supported Collaborative Learning* (2016). ISSN: 1556-1615. DOI: 10.1007/s11412-016-9235-5.

[28] W. Duch, R. J. Oentaryo, and M. Pasquier. “Cognitive Architectures: Where do we go from here?” In: *AGI*. Vol. 171. 2008, pp. 122–136.

[29] G. Durantin, S. Heath, and J. Wiles. “Social Moments: A Perspective on Interaction for Social Robotics”. In: *Frontiers in Robotics and AI* 4 (June 2017). DOI: 10.3389/frobt.2017.00024. URL: https://doi.org/10.3389/frobt.2017.00024.

[30] *Ethics Guidelines for Trustworthy AI*. High-level Expert Group on Artificial Intelligence, European Union, 2019. DOI: 10.2759/346720. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[31] J. Flavell, H. Beilin, and P. Pufall. “Perspectives on perspective taking”. In: *Piaget’s theory: Prospects and possibilities* (1992), pp. 107–139.

[32] U. Frith and F. Happé. “Autism: Beyond “theory of mind””. In: *Cognition* 50.1 (1994), pp. 115–132.

[33] I. Garcìa-Magariño, C. Medrano, A. S. Lombas, and A. Barrasa. “A hybrid approach with agent-based simulation and clustering for sociograms”. In: *Information Sciences* 345 (2016), pp. 81–95.

[34] S. S. Ge, Y. Yang, and T. H. Lee. “Hand gesture recognition and tracking based on distributed locally linear embedding”. In: *Image and Vision Computing* 26.12 (2008), pp. 1607–1620.

[35] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman. “Video action transformer network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 244–253.

[36] M. M. de Graaf, S. B. Allouch, and J. A. van Dijk. “A phased framework for long-term user acceptance of interactive technology in domestic environments”. In: *New Media & Society* 20.7 (Oct. 2017), pp. 2582–2603. DOI: 10.1177/1461444817727264. URL: https://doi.org/10.1177/1461444817727264.

[37] G. R. Greher, A. Hillier, M. Dougherty, and N. Poto. “SoundScape: An Interdisciplinary Music Intervention for Adolescents and Young Adults on the Autism Spectrum.” In: *International Journal of Education & the Arts* 11.9 (2010), n9.

[38] H. Gunes and B. Schüller. “Automatic Analysis of Social Emotions”. In: Cambridge University Press, 2017, p. 213. DOI: 10.1017/9781316676202.016.

[39] F. Happé. “Communicative competence and theory of mind in autism: A test of relevance theory”. In: *Cognition* 48.2 (1993), pp. 101–119.

[40] N. Hawes, C. Burbridge, F. Jovan, L. Kunze, B. Lacerda, L. Mudrova, J. Young, J. Wyatt, D. Hebesberger, T. Kortner, et al. “The strands project: Long-term autonomy in everyday environments”. In: *IEEE Robotics & Automation Magazine* 24.3 (2017), pp. 146–156.

[41] P. Heikkilä, H. Lammi, and K. Belhassein. “Where Can I Find a Pharmacy?: Human-Driven Design of a Service Robot’s Guidance Behaviour”. In: *4th Workshop on Public Space Human-Robot Interaction, PubRob 2018: Held as part of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2018)*. 2018.

[42] G. Hoffman. “Anki, Jibo, and Kuri: What We Can Learn from Social Robots That Didn’t Make It”. In: *IEEE Spectrum* (May 2019). URL: https://spectrum.ieee.org/automaton/robotics/home-robots/anki-jibo-and-kuri-what-we-can-learn-from-social-robotics-failures.

[43] F. Ingrand, R. Chatila, R. Alami, and F. Robert. “PRS: A high level supervision and control language for autonomous mobile robots”. In: *Proceedings of the IEEE International Conference on Robotics and Automation*. Vol. 1. 1996, pp. 43–49.

[44] H. Jaeger. “Controlling recurrent neural networks by conceptors”. In: *arXiv preprint arXiv:1403.3369*. Jacobs University Technical Reports 31 (2014).

[45] P. Jermann, G. Zufferey, B. Schneider, A. Lucci, S. Lépine, and P. Dillenbourg. “Physical space and division of labor around a tabletop tangible simulation”. In: *Proceedings of the 9th international conference on Computer supported collaborative learning-Volume 1*. 2009, pp. 345–349.

[46] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. *Children speech recording*. Dec. 2016. DOI: 10.5281/zenodo.200495.

[47] R. Kingdon. *A review of cognitive architectures*. Tech. rep. ISO Project Report. MAC 2008-9, 2008.

[48] L. Kunze, N. Hawes, T. Duckett, M. Hanheide, and T. Krajník. “Artificial intelligence for long-term robot autonomy: a survey”. In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4023–4030.

[49] R. Lallement, L. De Silva, and R. Alami. “HATP: An HTN Planner for Robotics”. In: *Proceedings of the PlanRob 2014, ICAPS*. 2014.

[50] P. Langley, J. E. Laird, and S. Rogers. “Cognitive architectures: Research issues and challenges”. In: *Cognitive Systems Research* 10.2 (2009), pp. 141–160.

[51] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. “Empathic Robots for Long-term Interaction”. In: *International Journal of Social Robotics* 6.3 (Mar. 2014), pp. 329–341. DOI: 10.1007/s12369-014-0227-1. URL: https://doi.org/10.1007/s12369-014-0227-1.

[52] I. Leite, C. Martinho, and A. Paiva. “Social Robots for Long-Term Interaction: A Survey”. In: *International Journal of Social Robotics* 5.2 (Apr. 2013), pp. 291–308. ISSN: 1875-4805. DOI: 10.1007/s12369-013-0178-y. URL: https://doi.org/10.1007/s12369-013-0178-y.

[53] S. Lemaignan and P. Dillenbourg. “Mutual Modelling in Robotics: Inspirations for the Next Steps”. In: *Proceedings of the 2015 ACM/IEEE Human-Robot Interaction Conference*. 2015.

[54] S. Lemaignan, C. Edmunds, and T. Belpaeme. *The PInSoRo dataset*. Dec. 2018. DOI: 10.5281/zenodo.1043507.

[55] S. Lemaignan, J. Fink, P. Dillenbourg, and C. Braboszcz. “The Cognitive Correlates of Anthropomorphism”. In: *Proceedings of the Workshop: A bridge between Robotics and Neuroscience at the 2014 ACM/IEEE Human-Robot Interaction Conference*. 2014.

[56] S. Lemaignan, M. Gharbi, J. Mainprice, M. Herrb, and R. Alami. “Roboscopie: A Theatre Performance for a Human and a Robot”. In: *Proceedings of the 2012 ACM/IEEE Human-Robot Interaction Conference*. 2012.

[57] S. Lemaignan, A. Jacq, D. Hood, F. Garcia, A. Paiva, and P. Dillenbourg. “Learning by Teaching a Robot: The Case of Handwriting”. In: *IEEE Robotics and Automation Magazine* (2016).

[58] S. Lemaignan, R. Ros, L. Mösenlechner, R. Alami, and M. Beetz. “ORO, a knowledge management module for cognitive architectures in robotics”. In: *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2010. DOI: 10.1109/IROS.2010.5649547.

[59] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami. “Artificial Cognition for Social Human-Robot Interaction: An Implementation”. In: *Artificial Intelligence* (2017). DOI: 10.1016/j.artint.2016.07.002.

[60] S. Lemaignan, C. E. R. Edmunds, E. Senft, and T. Belpaeme. “The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics”. In: *PLOS ONE* 13.10 (Oct. 2018), pp. 1–19. DOI: 10.1371/journal.pone.0205999. URL: https://doi.org/10.1371/journal.pone.0205999.

[61] S. Lemaignan, R. Ros, E. A. Sisbot, R. Alami, and M. Beetz. “Grounding the Interaction: Anchoring Situated Discourse in Everyday Human-Robot Interaction”. In: *International Journal of Social Robotics* (2011), pp. 1–19. ISSN: 1875-4791. URL: http://dx.doi.org/10.1007/s12369-011-0123-x.

[62] S. Lemaignan, Y. Sallami, C. Wallbridge, A. Clodic, and R. Alami. “underworlds: Cascading Situation Assessment for Robots”. In: *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2018. DOI: 10.1109/IROS.2018.8594094.

[63] A. Leslie and L. Thaiss. “Domain specificity in conceptual development: Neuropsychological evidence from autism”. In: *Cognition* 43.3 (1992), pp. 225–251.

[64] M. G. Madden and T. Howley. “Transfer of experience between reinforcement learning environments with progressive difficulty”. In: *Artificial Intelligence Review* 21.3-4 (2004), pp. 375–398.

[65] J. Mainprice, E. A. Sisbot, L. Jaillet, J. Cortes, R. Alami, and T. Simeon. “Planning human-aware motions using a sampling-based costmap planner”. In: *IEEE International Conference on Robotics and Automation*. 2011.

[66] M. Marmpena, A. Lim, T. S. Dahl, and N. Hemion. “Generating robotic emotional body language with variational autoencoders”. In: *8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019, pp. 545–551. DOI: 10.1109/ACII.2019.8925459.

[67] P. Marshall, Y. Rogers, and N. Pantidi. “Using F-formations to analyse spatial patterns of interaction in physical environments”. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. 2011, pp. 445–454.

[68] N. Martelaro, V. C. Nneji, W. Ju, and P. Hinds. “Tell Me More: Designing HRI to Encourage More Trust, Disclosure, and Companionship”. In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. HRI ’16. Christchurch, New Zealand: IEEE Press, 2016, pp. 181–188. ISBN: 978-1-4673-8370-7. URL: http://dl.acm.org/citation.cfm?id=2906831.2906863.

[69] R. Martinez-Maldonado, J. Kay, S. Buckingham Shum, and K. Yacef. “Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data”. In: *Human–Computer Interaction* 34.1 (2019), pp. 1–50.

[70] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013), pp. 3111–3119.

[71] G. Milliez, R. Lallement, M. Fiore, and R. Alami. “Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring”. In: *The Eleventh ACM/IEEE International Conference on Human Robot Interation*. IEEE Press. 2016, pp. 43–50.

[72] Y. Mohamed and S. Lemaignan. *ROS for Human-Robot Interaction*. 2020. arXiv: 2012.13944 [cs.RO].

[73] C. Moulin-Frier, S. M. Nguyen, and P.-Y. Oudeyer. "Self-organization of early vocal development in infants and machines: the role of intrinsic motivation". In: *Frontiers in Psychology* 4 (2014), p. 1006. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.01006. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2013.01006.

[74] P.-Y. Oudeyer, F. Kaplan, V. V. Hafner, and A. Whyte. "The playground experiment: Task-independent development of a curious robot". In: *Proceedings of the AAAI Spring Symposium on Developmental Robotics*. Stanford, California. 2005, pp. 42–47.

[75] A. K. Pandey and R. Alami. "Affordance graph: A framework to encode perspective taking and effort based affordances for day-to-day human-robot interaction". In: *IROS 2013, IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2013, pp. 2180–2187.

[76] A. Parmiggiani, L. Fiorio, A. Scalzo, A. V. Sureshbabu, M. Randazzo, M. Maggiali, U. Pattacini, H. Lehmann, V. Tikhanoff, D. Domenichelli, et al. "The design and validation of the R1 personal humanoid". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 674–680.

[77] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. "Curiosity-driven exploration by self-supervised prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 16–17.

[78] J. Pennington, R. Socher, and C. D. Manning. "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, pp. 1532–1543.

[79] J. Perner, U. Frith, A. Leslie, and S. Leekam. "Exploration of the autistic child's theory of mind: Knowledge, belief, and communication". In: *Child development* (1989), pp. 689–700.

[80] A. Saffiotti and M. Broxvall. "PEIS ecologies: Ambient intelligence meets autonomous robotics". In: *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies*. ACM. 2005, pp. 277–281.

[81] Y. Sallami, S. Lemaignan, A. Clodic, and R. Alami. "Simulation-based physics reasoning for consistent scene estimation in an HRI context". In: *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2019. DOI: 10.1109/IROS40897.2019.8968106.

[82] F. Schroff, D. Kalenichenko, and J. Philbin. "Facenet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.

[83] E. Senft, P. Baxter, J. Kennedy, S. Lemaignan, and T. Belpaeme. "Supervised Autonomy for Online Learning in Human-Robot Interaction". In: *Pattern Recognition Letters* (2017). DOI: 10.1016/j.patrec.2017.03.015.

[84] E. Senft, S. Lemaignan, P. Baxter, M. Bartlett, and T. Belpaeme. "Teaching robots social autonomy from in situ human guidance". In: *Science Robotics* (2019). DOI: 10.1126/scirobotics.aat1186.

[85] A. Sharkey and N. Sharkey. “Granny and the robots: ethical issues in robot care for the elderly”. In: *Ethics and information technology* 14.1 (2012), pp. 27–40.

[86] M. Shiomi, A. Nakata, M. Kanbara, and N. Hagita. “A Robot that Encourages Self-disclosure by Hug”. In: *Social Robotics*. Ed. by A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He. Cham: Springer International Publishing, 2017, pp. 324–333. ISBN: 978-3-319-70022-9.

[87] E. A. Sisbot, A. Clodic, R. Alami, and M. Ransan. “Supervision and Motion Planning for a Mobile Manipulator Interacting with Humans”. In: *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. 2008.

[88] B. Sodian and U. Frith. “Deception and sabotage in autistic, retarded and normal children”. In: *Journal of Child Psychology and Psychiatry* 33.3 (1992), pp. 591–605.

[89] B. C. Stahl and M. Coeckelbergh. “Ethics of healthcare robotics: Towards responsible research and innovation”. In: *Robotics and Autonomous Systems* 86 (2016), pp. 152–161. ISSN: 0921-8890. DOI: https://doi.org/10.1016/j.robot.2016.08.018. URL: http://www.sciencedirect.com/science/article/pii/S0921889016305292.

[90] M. Suguitan and G. Hoffman. “MoveAE: Modifying Affective Robot Movements Using Classifying Variational Autoencoders”. In: *Proceedings of the 2020 ACM/IEEE Human-Robot Interaction Conference*. 2020. DOI: 10.1145/3319502.3374807.

[91] N. Taatgen and J. R. Anderson. “The past, present, and future of cognitive architectures”. In: *Topics in Cognitive Science* 2.4 (2010), pp. 693–704.

[92] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita. “Perceiving the person and their interactions with the others for social robotics–a review”. In: *Pattern Recognition Letters* 118 (2019), pp. 3–13.

[93] M. Tenorth and M. Beetz. “KnowRob – knowledge processing for autonomous personal robots”. In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, pp. 4261–4266.

[94] K. R. Thórisson and H. P. Helgasson. “Cognitive Architectures and Autonomy: A Comparative”. In: *Journal of Artificial General Intelligence* 3.2 (2012), pp. 1–30.

[95] G. Trafton, L. Hiatt, A. Harrison, F. Tamborello, S. Khemlani, and A. Schultz. “ACT-R/E: An embodied cognitive architecture for human-robot interaction”. In: *Journal of Human-Robot Interaction* 2.1 (2013), pp. 30–55.

[96] R. Triebel, K. Arras, R. Alami, L. Beyer, S. Breuers, R. Chatila, M. Chetouani, D. Cremers, V. Evers, M. Fiore, et al. “Spencer: A socially aware service robot for passenger guidance and help in busy airports”. In: *Field and service robotics*. Springer. 2016, pp. 607–622.

[97] S. Tulli, D. A. Ambrossio, A. Najjar, and F. J. R. Lera. “Great Expectations & Aborted Business Initiatives: The Paradox of Social Robot Between Research and Industry”. In: *Proceedings of the Reference AI & ML Conference for Belgium, Netherlands & Luxemburg*. 2019.

[98] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[99] D. Vernon, G. Metta, and G. Sandini. “A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents”. In: *IEEE Transactions on Evolutionary Computation* 11.2 (2007), p. 151.

[100] J. M. K. Westlund, H. W. Park, R. Williams, and C. Breazeal. “Measuring children’s long-term relationships with social robots”. In: *Workshop on Perception and Interaction dynamics in Child-Robot Interaction, held in conjunction with the Robotics: Science and Systems XIII*. 2017.

[101] M.-A. Williams. *Social Robotics*. Jan. 2020. URL: https://www.xplainableai.org/socialrobotics/.

[102] A. F. Winfield and M. Jirotka. “The case for an ethical black box”. In: *Annual Conference Towards Autonomous Robotic Systems*. Springer. 2017, pp. 262–273.

[103] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. “Couch to 5km Robot Coach: An Autonomous, Human-Trained Socially Assistive Robot”. In: *Companion Proceedings of the 2020 ACM/IEEE Human-Robot Interaction Conference*. 2020. DOI: 10.1145/3371382.3378337.

[104] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. “In-Situ Learning from a Domain Expert for Real World Socially Assistive Robot Deployment”. In: *Proceedings of Robotics: Science and Systems 2020*. 2020. DOI: 10.15607/RSS.2020.XVI.059.

[105] K. Winkle, P. Caleb-Solly, A. Turton, and P. Bremner. “Social Robots for Engagement in Rehabilitative Therapies: Design Implications from a Study with Therapists”. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI ’18. New York, NY, USA: ACM, 2018, pp. 289–297. ISBN: 978-1-4503-4953-6. DOI: 10.1145/3171221.3171273.

[106] F. Yang and C. Peters. “AppGAN: Generative Adversarial Networks for Generating Robot Approach Behaviors into Small Groups of People”. In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2019, pp. 1–8. DOI: 10.1109/RO-MAN46459.2019.8956425.

[107] G.-Z. Yang, J. Bellingham, P. E. Dupont, P. Fischer, L. Floridi, R. Full, N. Jacobstein, V. Kumar, M. McNutt, R. Merrifield, et al. “The grand challenges of Science Robotics”. In: *Science robotics* 3.14 (2018), eaar7650.