CNRS Research Project

Socially-Driven Autonomous Robots

for

Real-world Human-Robot Interactions

Pr. Séverin Lemaignan

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 years a socially-intelligent and responsible robot, that (1) will have recognised social utility, and (2) will see long-term acceptance by its 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 driven by the *intention* to support positive human-human interactions. How this general principle translates into specific guidelines and algorithms – while taking into account the principles of a responsible AI – is a central contribution of the 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 the robot how to become a good and responsible social helper.

My research tests these two hypotheses with an ambitious 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 project opens a unique window into the positive role social robots can play in our future societies; it will provide a lasting legacy, paving the way forward for a better understanding of the design of socially-intelligent robots that are socially useful and acceptable in the long-term.

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.

The Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS), part of the Artificial and Natural Intelligence Toulouse Institute (ANITI), would be an ideal host laboratory to successfully conduct this programme: its strong track-record in autonomous interactive robots, combined with the breadth of expertise available within the ANITI institute, would prove instrumental in scaffolding and accelerating several of the key science breakthrough I target with this project.

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, Fink, 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 (G.-Z. 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. (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, Martinho, et al. 2013), with a number of studies involving social robots deployed in real-world settings (for instance in schools (Leite, Castellano, et al. 2014; Westlund et al. 2017; Lemaignan, Jacq, et al. 2016; Coninx et al. 2016), homes (Graaf et al. 2017) and care centres (Hawes et al. 2017; Winkle, Lemaignan, 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 (Triebel et al. 2016) projects), or a narrow, well-specified complex task (for instance, supporting exercising in a gym, as I did in (Winkle, Lemaignan, et al. 2020a)). However, general purpose, long-term interaction is still an open question.

Novelty, context, timeliness, relevance

The service and companion robots that we are set to interact with in the coming years are being designed and built today in labs and startups all over the world. Indeed, we already envision close and long-term human-robot interactions in a range of sensitive domains like education, elderly care and health care. Critically, we as a society, need to develop in parallel the underpinning principles that will ensure the future roles of social robots are collectively defined, in a responsible and ethical manner – in particular in the context their interactions with vulnerable populations.

Progressing this question requires real-world evidence. However, because autonomous social robots lie at the forward edge of science and engineering, the real-world, long-term deployments required to gather such evidence are extremely rare. As a consequence, we currently have limited insights into the factors that determine the utility and acceptability of social robots.

My research programme approaches this important and timely question in a **novel and ambitious** manner: the project will define and implement **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 to create a **state-of-the-art autonomous social robot** that not only learns social behaviours **with and from** the public and end-users, but is also **co-designed from the ground-up to be acceptable, responsible and useful** to the humans it will serve.

Ambition, adventure, transformative aspects

This research is ground-breaking: **My programme will lead to 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**.

The project is **ambitious**: in the next 5 to 10 years, I will have brought together two emerging AI paradigms (teleological architectures and human-in-the-loop machine learning); I will have them integrated into a state-of-the-art cognitive architecture for autonomous social robots, relying on multidisciplinary approaches where relevant (eg. a choreographer to create a novel 'body language' for social robots); I will have created the conditions for a unique, large-scale, 'public-in-the-loop' participatory design approach that will transform how we think about public engagement with technology design; finally, I will have co-designed and deployed autonomous robots in several real-world and highly social settings, for significant periods of time.

Combining scientific ambition, engineering ambition and methodological ambition, my research programme sets a high bar for excellence. Surprisingly few groups worldwide have achieved full autonomy for a complex social robot – the LAAS being one of them. By joining the laboratory, I would create the conditions to 'future-proof' this scientific know-how, while developing a wide-ranging set of new research directions that promise to have a transformative impact on our digital future.

Methodology and approach to achieve impact

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:

RQ1: What are the public expectations with respect to the role of social robots, and how can we collectively design principles ensuring responsible, beneficial, socially acceptable robots?

RQ2: What are the conceptual, algorithmic and technical prerequisites to design and implement such an embodied AI? in particular, what AI is required to **sustain long-term engagement** between end-users and a robot?

RQ3: 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 guidelines.

O2: real–time social modeling To create the novel cognitive capability of artificial *social situation assessment* and enabe the robot to represent real-time social dynamics in its environment, I will significantly extend and integrate the current state-of-art in spatio-temporal modeling (so-called *situation assessment*) with my recent research in social state modeling.

O3: congruent social behaviours production To create a novel way of producing non-repetitive, socially-congruent, expressive motions using the state of the art in generative neural networks, combined with data acquired from an expert choreographer. This will be integrated with novel *sound landscapes* to create a beyond-state-of-art, non-verbal (yet highly expressive) action and communication system for the robot.

O4: embodied AI breakthrough 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 and already used in complex social environments (Senft, Baxter, et al. 2017; Senft, Lemaignan, Baxter, Bartlett, et al. 2019; Winkle, Lemaignan, 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 (Senft, Lemaignan, Baxter, Bartlett, et al. 2019), this approach leads to increased control and ownership of the system, and as a result, increased trust on the part of 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.

Implementation of the work programme

The five scientific objectives presented in the previous section 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;
- **Strand 2** investigates the **intelligent generation of social behaviours**, exploring novel techniques mixing immersive teleoperation and adversarial generative networks;
- **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.
- **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**.

Those four research strands are all underpinned by one additional research activity, the transversal **Strand 5**. As a research scientist, I will build-up an ambitious **experimental capacity** at LAAS-CNRS, that will significantly improve upon the current, mostly lab-based, experimental work conducted to date. Building on my extensive experience in real-world deployments of autonomous social robots (see section **??**), 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).

Add potential collaborations to each research strands, ANITI + non-ANITI

Overview and coherence of the research programme

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, Toulouse's Purpan hospital, Toulouse's science centre 'Le Quai des Savoirs', etc) will be integral to my research methodology, and will en-

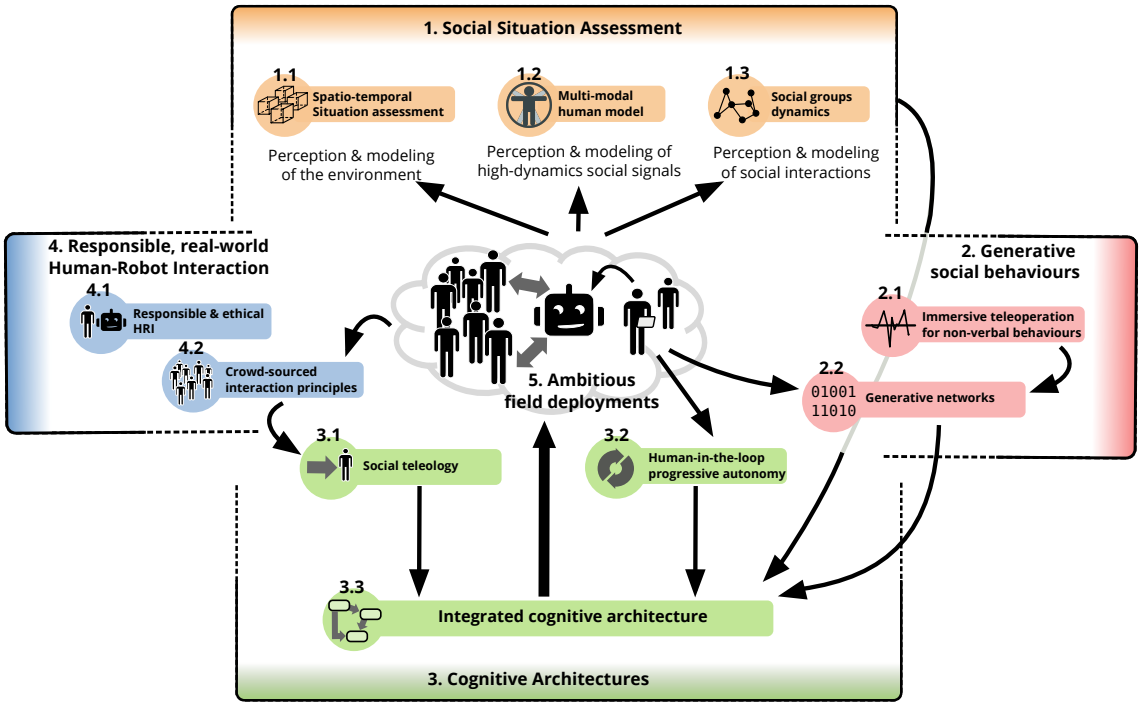

Figure 1: Overview of the research strands that I intend to develop as a CNRS research scientist.

able the development of 'public-in-the-loop' experiments (4.2): 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 to 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 in the following sections.

These perceptual capabilities are both (1) continuously integrated into the robot's cognitive architecture (3.3), iteratively improving the socio-cognitive performances of the robot, (2) disseminated to the broader community through standardisation and integration to the 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.

My research project includes an ambitious, beyond-state-of-art, technical work programme, and **Strand 3** investigates the *principled* integration of a cognitive architecture for autonomous social robots. Indeed, in addition to the integration of the results of Strand 1 and Strand 2, Strand 3 is also 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.

Towards building a principled cognitive architecture

As such, an important part of my research programme contributes to the design and implementation of a novel, complete cognitive architecture for socially intelligent robots. While I detail in a following section the scientific underpinnings of this research, Figure 2 illustrates how my different research strands combine towards that goal.

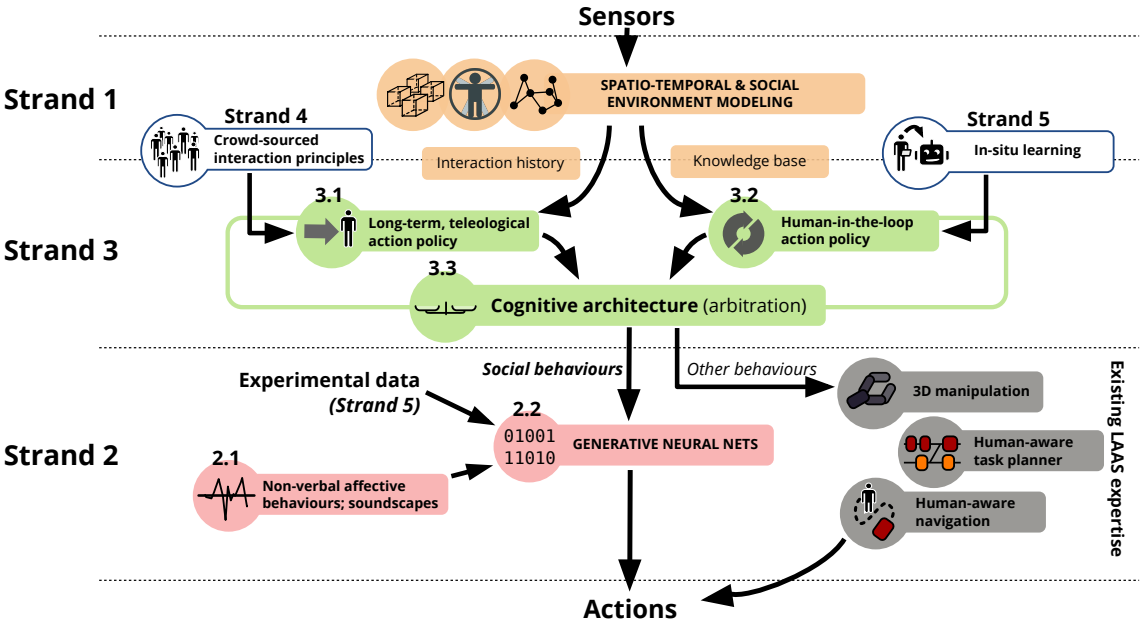

Figure 2: Contributions of my research strands to the AI architecture at the core of the research programme. Some capabilities already developed at the host lab (grey blocks) will directly contribute to the practical realisation of autonomy.

Strand 1 (top) focuses on creating a novel, integrated model of the social environment of the robot; it will build on the current state of art in spatial modeling, semantic modeling and interaction history representation, and augment it with representations of the social dynamics around the robot, introducing the idea of *social embeddings*. Strand 2 (bottom) significantly improve upon techniques for non-repetitive, socially-congruent behaviour production, combining behaviour design and data acquisition using immersive teleoperation with recent advances in generative neural nets. Strand 3 (centre) integrates the robot cognitive capabilities in a new cognitive architecture for long-term social autonomy. It introduces a novel arbitration mechanism between action policies, to enable both long-term, goal-driven

autonomous behaviours, and direct in-situ learning from the robot's end-users, to ensure transparency and human oversight.

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

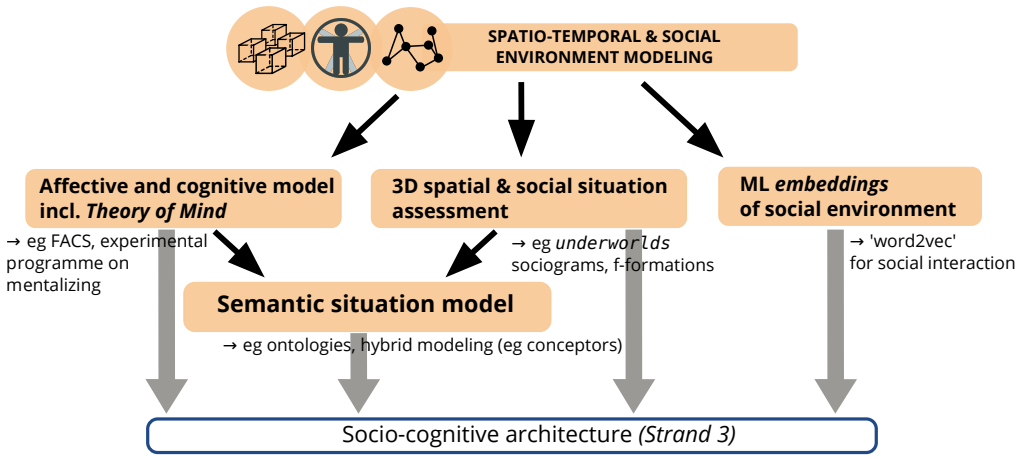

Figure 3: Social situation assessment, from feature extraction to 3D modelling, to social *embeddings* and symbolic reasoning. I will focus on using ROS for implementation, facilitating broad and lasting impact on the community.

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 3).

Indeed, knowledge representation and grounding is a fundamental building block for cognitive architectures, as I have shown (Lemaignan, Warnier, et al. 2017), in fact this first research strand builds on existing work on symbolic knowledge representation (eg. (Tenorth and Beetz 2009) or my own work (Lemaignan, Ros, Mösenlechner, et al. 2010)) and my work on situation assessment (Lemaignan, Sallami, et al. 2018; 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

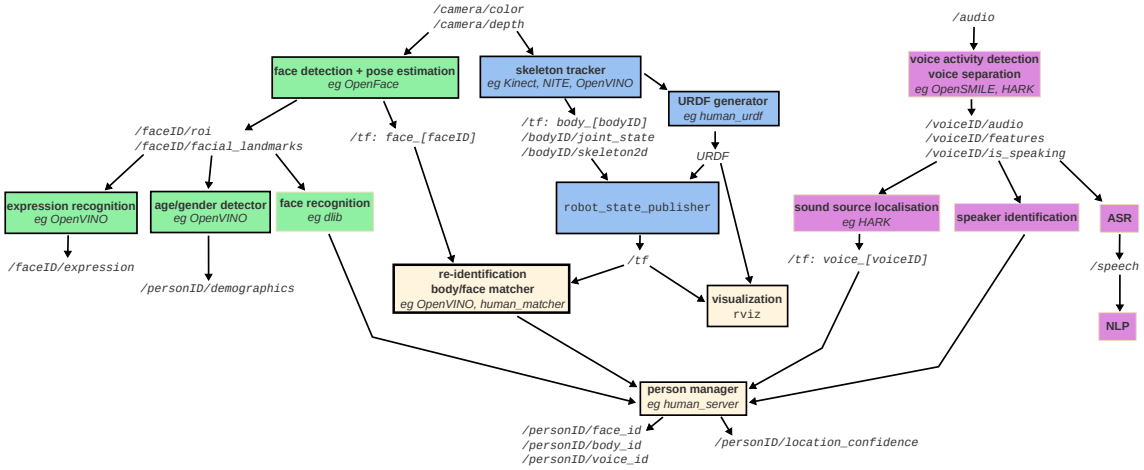

Figure 4: Overview of the 'ROS4HRI' pipeline that I currently develop.

vicinity. I have recently introduced a dataset of social interaction (Lemaignan, Edmunds, et al. 2018) 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). This dataset and the related methodologies on data-driven social modeling will form the basis of this research workpackage, and will exploit the natural interaction data collected in Strand 5.

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 on social dynamics analysis (eg (Durantin et al. 2017; Jermann et al. 2009; Martinez-Maldonado et al. 2019)) to 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 and perspective taking (Flavell et al. 1992) (as I previously framed in (Lemaignan and Dillenbourg 2015; Dillenbourg et al. 2016)).

Research strand 1.2 – targeted outcomes:

A novel cognitive sub-system for social situation assessment, released as an open-source set of integrated ROS modules ("*ROS4HRI*", Figure 4). This set of tools 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.

1.3 – Social embeddings

One of the key novel scientific idea that I will research in this workpackage is the construction of the **social embedding** of the robot: a compact, low-dimensional representation of the full social environment, inspired from word embeddings (eg. (Mikolov et al. 2013)). 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 (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.

Experimental programme of Strand 1

In complement to the large-scale experimental work described in Strand 5, a focused experimental programme accompanies Strand 1, to demonstrate (in relative isolation) the resulting socio-cognitive capabilities. I will implement a subset of the experimental protocols identified by Frith and Happé (Frith and F. 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. The Sally and Anne task, for instance, requires the observing child to be able to visually follow the marble, to remember the true location of the marble, to understand simple questions ("Where will Sally look for her marble?" in Baron-Cohen's protocol (Baron-Cohen, A. Leslie, et al. 1985)) and eventually to give an answer, either verbally or with a gesture – the two first points being actually explicitly checked through questions: "Where is the marble really?" (reality control question) and "Where was the marble in the beginning?" (memory control question).

Likewise, current social robots have limited cognitive skills (no fast yet fine motor skills, limited speech production and understanding, limited scene segmentation and object recognition capabilities, etc.) and such tasks that effectively test a single cognitive skill (in this case, mentalizing) in near isolation are of high relevance for experimental social robotics (Lemaignan and Dillenbourg 2015), offering rich experimental opportunities, early on in the development process of the complete cognitive architecture.

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 men-

No mentalizing required	Mentalizing required
Ordering behavioural pictures	Ordering mentalistic pictures(Baron-Cohen, A. M. Leslie, 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(A. M. Leslie and Thaiss 1992)
Recognizing happiness and sadness	Recognizing surprise(Baron-Cohen, Spitz, et al. 1993)
Object occlusion	Information occlusion(Baron-Cohen 1992)
Literal expression	Metaphorical expression(F. G. Happé 1993)

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

talizing), thus providing control tasks. *Object occlusion* vs. *Information occlusion* is one example of a pair of tasks which evidence representation-level perspective taking through *adaptive deception*. Adapting such a protocol for a human-robot pair 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 (however, see Section for planned collaborations on *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).

This strand of research will build upon the long expertise of the LAAS-CNRS on developing social robots interacting with humans in complex environments Lallement et al. 2014; Gharbi et al. 2013; Waldhart et al. 2015**add additional/better references**, as well as the existing literature on current behaviour generation methodologies, covering techniques like curiosity-driven behaviours(Oudeyer et al. 2005), Learning from Demonstration(Billard et al. 2008; Argall et al. 2009), human-in-the-loop action policy learning(Senft, Lemaignan, Baxter, and Belpaeme 2016; Senft, Lemaignan, Baxter, Bartlett, et al. 2019).

Collaborating with Nicholas Asher, IRIT, ANITI

Dr. Nicholas Asher is a leading figure in natural language processing. Collabo-

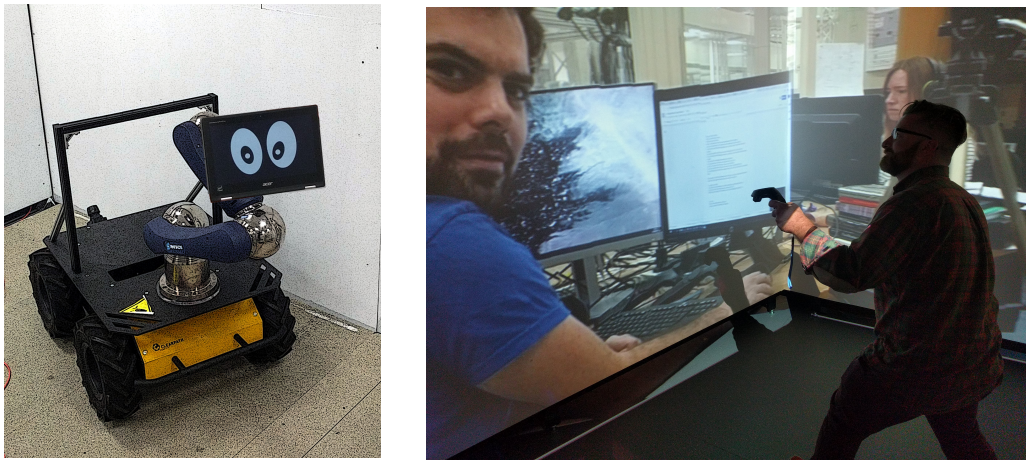

Figure 5: (left) Possible appearance of a puppet-robot that I will use to collect data. A tablet, displaying facial animations, is mounted on a robotic arm. It can freely orient its 'gaze' and use expressive movements. The robot is effectively teleoperated in realtime by an artist (eg choreographer, right) who 'sees through the robots' eyes' (Bailly et al. 2015) and who is tasked with designing and acting a novel 'body language' for social interactions.

rating with him would open strong mutual opportunities to develop and apply language processing for physical, embodied agents. the LAAS would enable me to

2.1 – Immersive teleoperation to design richer non-verbal interactions

As part of Strand 2, I also intend to lead research on novel non-verbal interaction modality for social robots. This direction of research, tightly related to the previous one, will pursue an interdisciplinary approach: from creating a novel body language for social robot with choreographers, to investigating new forms of sound expressions like soundscapes with sound experts: soundscapes are about creating a sound environment that reflects a particular situation; they also have been shown to be an effective intervention technique in the context special needs treatments (eg (Greher et al. 2010)). The soundscapes that we will research and create, are 'owned' by the robot, and it can manipulate it itself, eg to create an approachable, non-threatening, non-judgmental, social interaction context, or to the establish the interaction into a trusted physical and emotional safe-space for the children.

Complete the 'immersive teleoperation for behaviour design' part

with a dataset co-created with artists (for instance, a choreographer): during a period of time, a choreographer would join the lab and remotely 'puppet' a robot which would be itself interacting with the lab members (Figure 5)

Collaborating with Olivier Stasse, LAAS-CNRS

Olivier Stasse is an expert in humanoid motion, and has past experience in working with dancers to inform the design of complex motions. I have a track-record of collaboration with him on art-driven design methodologies (Lemaignan, Gharbi, et al. 2012).

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).

The LAAS-CNRS has played a pioneering role in this field with eg the development of HATP as a hybrid task-planner for human-robot interaction Alili et al. 2008; Lallement et al. 2014 or human-aware motion planning Gharbi et al. 2013; Waldhart et al. 2015. 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 (F. 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, Lemaignan, Baxter, Bartlett, et al. 2019; Winkle, Lemaignan, et al. 2020b) that fully involve the end-users (Winkle, Caleb-Solly, 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, Lemaignan, Baxter, Bartlett, et al. 2019; Winkle, Lemaignan, 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, Baxter, 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, Warnier, 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 cognitive architecture, implemented on the LAAS social robots, 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 human–robot 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 (Artificial Intelligence 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 (A. Sharkey and N. 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 sensitive social spaces

I have the ambition to demonstrate long-term, co-designed social interactions in two complex, socially sensitive spaces.

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 one year, in close partnership with local institutions (like Toulouse's "Le Quai des Savoirs"), the general public will be invited to teleoperate my robots, with the objective of interacting and assisting other visitors. The participants will remotely control the robot through a tablet interface (similar to the setup I created for (Senft, Lemaignan, Baxter, Bartlett, et al. 2019) and (Winkle, Lemaignan, et al. 2020b)), and interviews of both the teleoperators and the visitors interacting with the robot will be conducted in parallel, collecting in a structured manner the interaction patterns and social norms that will emerge over the course of the study. Additional focus groups will be organised with the public to reflect and iterate on these principles.

finish that During the duration of the study, one researcher will be permanently based at the science museum, and the museum staff themselves will be trained to communicate about the aims of the study. Anonymous interaction data (eg, body postures) will be collected as well, and feed into WP2 and WP3.

Collaborating with Toulouse Science Centre *Le Quai des Savoirs* & local science charities

This work will only be successful if it reaches out to a broad and diverse public. I have an established network of contact in the local science communication community (eg local science charities *Planète Sciences*, *Les Petits Débrouillards*) and intend to establish a solid collaboration with Toulouse Science Centre *Le Quai des Savoirs* where field studies could be run.

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

I intend to create a strong culture of Human-Robot Interaction experimental work at the LAAS-CNRS: 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, Baxter, 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.

do I re-introduce a risk assessment section?

Adequation of the host laboratory to the research project

The **Laboratoire d'Analyse et d'Architecture des Systèmes** (LAAS), part of the **Artificial and Natural Intelligence Toulouse Institute** (ANITI), would be an ideal host laboratory to successfully conduct my research programme: its strong track-record in autonomous interactive robots, combined with the breadth of expertise available within the ANITI institute, would prove instrumental in scaffolding and accelerating several of the key science breakthrough I target with this project.

LAAS expertise

The LAAS is one of the few research group worldwide that has achieved full autonomy for complex socio-cognitive robots (eg Lemaignan, Warnier, et al. 2017).

This has been made possible by the long-term commitment of the laboratory to cognitive robotics, spearheaded by Pr. Rachid Alami. Since the mid-90's, Alami and colleagues have indeed build a large, ambitious body of research, enabling a wide range of cognitive skills on service robots. To cite a few: reasoning languages Ingrand et al. 1996; symbolic reasoning Lemaignan, Ros, Mösenlechner, et al. 2010 and supervision Clodic et al. 2009; 3D motion and task planning Sisbot et al. 2008; Mainprice et al. 2011; human-aware symbolic task planning Alili et al. 2008; Lallement et al. 2014; Milliez et al. 2016; cognitive architectures Lemaignan, Warnier, et al. 2017; Devin and Alami 2016; language understanding Lemaignan, Ros, Sisbot, et al. 2011. These achievements are underpinned by excellent science, and also the strong commitment to technical excellence (eg Mallet et al. 2010), an often overlooked yet unique strength of the LAAS.

I took part myself to this scientific journey, from 2008 to 2012, contributing key cognitive elements (symbolic reasoning, perspective taking, language understanding), as well as integrating the research into a meaningful, coherent architecture Lemaignan, Warnier, et al. 2017.

These four years spent at LAAS, followed by 8 years in other institutions abroad put me in the unique position of both understanding in great detail and fully appreciating the approach and strengths of the LAAS, while also being able to appraise the potential weaknesses and direction of change.

I would like to outline here some

By joining the laboratory, I would create the conditions to 'future-proof' this scientific know-how, while developing a wide-ranging set of new research directions that promise to have a transformative impact on our digital future.

Integration within the broader local research landscape

TODO, in particular connections to ANITI Here, give a high-level overview of the potential for integration with existing structures; detailled collaboration opportunities to be specified earlier in the project.

- verbal behaviours -> collaboration within ANITI? - collaboration with Rafaelle
- experimental work: partnerships/institution in Toulouse

ANITI themes

ANITI, the *Artificial and Natural Intelligence Toulouse Institute* is one of the four French Institutes for AI, aiming at international leadership in AI. It brings together 200+ research scientists around Toulouse, and organises the research around three main programmes: Acceptability and AI; Certifiable AI; Collaborative AI.

My research programme aligns particularly well with these themes. I would directly benefit of excellent collaboration opportunities (see below), and in return, my programme would create unique transversal opportunities to tie together key sub-themes.

Responsible AI

- Fair & Robust Machine Learning https://aniti.univ-toulouse.fr/chaire-jean-michel-loubes/ Jean-Michel Loubès
- Law, Accountability and Social Trust in AI https://aniti.univ-toulouse.fr/chaire-celine-castets-renard/ Céline Castets-Renard
- Moral AI https://aniti.univ-toulouse.fr/chaire-moral-ai/ Jean-François Bonnefon
- New certification approaches of critical AI based systems https://aniti.univ-toulouse.fr/chaire-claire-pagetti/ Claire Pagetti

AI for policy learning

- Augmented Society https://aniti.univ-toulouse.fr/chaire-cesar-hidalgo/ Cesar Hidalgo
- Empowering Data-driven AI by Argumentation and Persuasion https://aniti.univ-toulouse.fr/chaire-leila-amgoud/ Leïla Amgoud

AI for complex behaviour generation

- Motion Generation for Complex Robots https://aniti.univ-toulouse.fr/chaire-nicolas-mansard/ Nicolas Mansard

Summary, importance and outlook

Refine concluding remarks

Summary of the research project

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.

Update that section for France/CNRS/Europe/ANITI

These questions are of prime societal importance, and this research closely aligns with the **EPSRC Delivery Plan *Connected Nation* and *Healthy Nation* priorities**. Specifically, the project investigates and will significantly advance the questions of Trustworthy autonomous AI, Multidisciplinary approaches to technology acceptability and Technology for the public good[1].

The project is also closely aligned with UKRI Healthcare Technology Grand Challenge: *Transforming Community Health and Care*[2] by significantly advancing our capabilities in term of socially assistive robotics.

More broadly, and as a multidisciplinary project, **change** relates to several themes of the EPSRC portfolio. The main ones are: *Human–computer interaction* and *Social computing/interactions* within the *Digital Economy* theme, *Assistive technology* within the *Healthcare technology* theme, and *Artificial Intelligence* and *Robotics* within the Engineering theme.

From an academic perspective, the UK 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 UK did play a key role in several of these projects (eg FP6-Cogniron, FP7-CHRIS, FP7-STRANDS, FP7-Poeticon++), and has built a solid reputation. It is now critical that this expertise is maintained and further developed, as to ensure the future academic leadership of the UK.

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 UK's strong societal values, the task T1.1, which continues throughout the project, will specifically address and frame 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

[1] EPSRC Delivery Plan 2019: https://epsrc.ukri.org/about/plans/dp2019/

[2] https://epsrc.ukri.org/research/ourportfolio/themes/healthcaretechnologies/strategy/grandchallenges/

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 psychosocial 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, intelligent robotics, learning technology. As such, **my research project builds bridges across multiple disciplinary boundaries**.

I deliver this programme by building on a range of multidisciplinary methods, including user-centered design; ethnographic and sociological investigation; expressive non-verbal communication, including 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 already been established in the Human-Robot Interaction group of LAAS-CNRS. I will seek funding to recruit PhDs and post-docs with 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 researcher 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 laboratory.

Outlook

This research project is ambitious, and I believe I am in a unique position to deliver on its work plan. I already have established international recognition in human-robot interaction and have likewise demonstrated strong leadership by leading research teams in three different institutions. As presented in my track-record, breadth of my interdisciplinary research covers the scientific expertise required by the project, providing me with a unique overall perspective and understanding of the domain. I am also a technology expert, with major software and hardware contributions to the robotic community. As such, I have a good grasp of the technical feasibility of the proposed work.

This research programme is indeed ambitious, with an experimental programme that goes significantly beyond the state of the art. It will provide a lasting scientific and technical legacy, and will inject a new momentum into the strong human-robot expertise of LAAS-CNRS. Finally, this research programme would also be a powerful enabler: it creates the opportunity to establish myself and the CNRS as world-leader in the emerging field of socially-driven, responsible autonomous robots, significantly reinforcing the national and European capacity in this critical field for our digital future.

References

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

[2] 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).

[3] L. Baillie et al. "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.

[4] S. Tulli et al. "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.

[5] 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.

[6] D. Dereshev et al. "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.

[7] 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.

[8] S. Lemaignan, J. Fink, et al. "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.

[9] G.-Z. Yang et al. "The grand challenges of Science Robotics". In: *Science robotics* 3.14 (2018), eaar7650.

[10] T. Belpaeme et al. "Social robots for education: A review". In: *Science robotics* 3.21 (2018), eaat5954.

[11] L. Kunze et al. "Artificial intelligence for long-term robot autonomy: a survey". In: *IEEE Robotics and Automation Letters* 3.4 (2018), pp. 4023–4030.

[12] 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.

[13] I. Leite, G. Castellano, et al. "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.

[14] J. M. K. Westlund et al. "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.

[15] S. Lemaignan, A. Jacq, et al. "Learning by Teaching a Robot: The Case of Handwriting". In: *IEEE Robotics and Automation Magazine* (2016).

[16] A. Coninx 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.

[17] N. Hawes et al. "The strands project: Long-term autonomy in everyday environments". In: *IEEE Robotics & Automation Magazine* 24.3 (2017), pp. 146–156.

[18] K. Winkle, S. Lemaignan, et al. "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.

[19] 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.

[20] R. Triebel 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.

[21] E. Senft, P. Baxter, et al. "Supervised Autonomy for Online Learning in Human-Robot Interaction". In: *Pattern Recognition Letters* (2017). DOI: 10.1016/j.patrec.2017.03.015.

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

[23] K. Winkle, S. Lemaignan, et al. "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.

[24] P. Baxter et al. "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.

[25] S. Lemaignan, M. Warnier, et al. "Artificial Cognition for Social Human-Robot Interaction: An Implementation". In: *Artificial Intelligence* (2017). DOI: 10.1016/j.artint.2016.07.002.

[26] 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.

[27] S. Lemaignan, R. Ros, L. Mösenlechner, et al. "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.

[28] S. Lemaignan, Y. Sallami, et al. "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.

[29] Y. Sallami et al. "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.

[30] 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.

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

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

[33] S. Lemaignan, C. E. R. Edmunds, et al. "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.

[34] M. Bartlett et al. "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.

[35] A. Tapus et al. "Perceiving the person and their interactions with the others for social robotics–a review". In: *Pattern Recognition Letters* 118 (2019), pp. 3–13.

[36] 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.

[37] I. Garcìa-Magariño et al. "A hybrid approach with agent-based simulation and clustering for sociograms". In: *Information Sciences* 345 (2016), pp. 81–95.

[38] 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.

[39] 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.

[40] P. Jermann et al. "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.

[41] R. Martinez-Maldonado et al. "Collocated collaboration analytics: Principles and dilemmas for mining multimodal interaction data". In: *Human–Computer Interaction* 34.1 (2019), pp. 1–50.

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

[43] 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.

[44] P. Dillenbourg et al. "The Symmetry of Partner Modelling". In: *Intl. J. of Computer-Supported Collaborative Learning* (2016). ISSN: 1556-1615. DOI: 10.1007/s11412-016-9235-5.

[45] T. Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26 (2013), pp. 3111–3119.

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

[47] S. Baron-Cohen, A. Leslie, and U. Frith. "Does the autistic child have a "theory of mind" ?" In: *Cognition* (1985).

[48] S. Baron-Cohen, A. M. 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.

[49] J. Perner et al. "Exploration of the autistic child's theory of mind: Knowledge, belief, and communication". In: *Child development* (1989), pp. 689–700.

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

[51] 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.

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

[53] 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.

[54] 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.

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

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

[57] M. Gharbi et al. "Natural Interaction for Object Hand-Over". In: *Proceedings of the 2013 ACM/IEEE Human-Robot Interaction Conference*. 2013.

[58] J. Waldhart, M. Gharbi, and R. Alami. "Planning handovers involving humans and robots in constrained environment". In: *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE. 2015, pp. 6473–6478.

[59] P.-Y. Oudeyer et al. "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.

[60] A. Billard et al. "Robot Programming by Demonstration". In: *Springer Handbook of Robotics*. Springer, 2008, pp. 1371–1394.

[61] B. D. Argall et al. "A Survey of Robot Learning From Demonstration". In: *Robotics and Autonomous Systems* 57.5 (2009), pp. 469–483.

[62] E. Senft, S. Lemaignan, P. Baxter, and T. Belpaeme. "SPARC: an efficient way to combine reinforcement learning and supervised autonomy". In: *Proc. of the Future of Interactive Learning Machines (FILM) Workshop, NIPS*. 2016.

[63] G. R. Greher et al. "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.

[64] 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.

[65] S. Lemaignan, M. Gharbi, et al. "Roboscopie: A Theatre Performance for a Human and a Robot". In: *Proceedings of the 2012 ACM/IEEE Human-Robot Interaction Conference*. 2012.

[66] 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.

[67] 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.

[68] M. Marmpena et al. "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.

[69] 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.

[70] K. Winkle, P. Caleb-Solly, et al. "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.

[71] 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.

[72] D. Pathak et al. "Curiosity-driven exploration by self-supervised prediction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 16–17.

[73] 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.

[74] 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.

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

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

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

[78] 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.

[79] 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.

[80] G. Trafton et al. "ACT-R/E: An embodied cognitive architecture for human-robot interaction". In: *Journal of Human-Robot Interaction* 2.1 (2013), pp. 30–55.

[81] 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.

[82] 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.

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

[84] 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.

[85] X. Chen et al. "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.

[86] A. Antunes et al. "From Human Instructions to Robot Actions: Formulation of Goals, Affordances and Probabilistic Planning". In: *ICRA*. 2016.

[87] 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.

[88] 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.

[89] H.-l. E. G. on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Tech. rep. European Commission, 2019. URL: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[90] 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.

[91] 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.

[92] N. Martelaro et al. "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.

[93] M. Shiomi et al. "A Robot that Encourages Self-disclosure by Hug". In: *Social Robotics*. Ed. by A. Kheddar et al. Cham: Springer International Publishing, 2017, pp. 324–333. ISBN: 978-3-319-70022-9.

[94] 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.

[95] C. Allen et al. *Robot ethics: the ethical and social implications of robotics*. MIT press, 2011.

[96] 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.

[97] F. Ingrand et al. "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.

[98] A. Clodic et al. "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.

[99] E. A. Sisbot et al. "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.

[100] J. Mainprice et al. "Planning human-aware motions using a sampling-based costmap planner". In: *IEEE International Conference on Robotics and Automation*. 2011.

[101] G. Milliez et al. "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.

[102] 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.

[103] S. Lemaignan, R. Ros, E. A. Sisbot, et al. "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.

[104] A. Mallet et al. "GenoM3: Building middleware-independent robotic components". In: *Proceedings of the 2010 IEEE International Conference on Robotics and Automation*. 2010.