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 is already part of our daily life, and robots are increasingly part of our everyday lives, supporting our ageing society, and 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 that, within 5 years, we create 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, with a combined two years of field deployment 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

Research project

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 (see *National and International Importance* section), this research project creates a unique opportunity to establish myself a key leader in Intelligent Social Robotics, as well as asserting the CNRS and European worldwide leadership in AI and 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 [105]. 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 [14]. 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 [5]. 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 [97].

Indeed, understanding *why* these robots have failed, is one of the active debate within the Human-Robot Interaction community [37], with only a handful of qualitative studies on this question [21, 32]. 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 [55]; the poor management of expectations, where user overascribe cognitive capabilities that do not match the reality. The community agrees however that the crux of the issue is achieving long-term social engagement [111, 37]

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. [10] 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 [46, 51], with a number of studies involving social robots deployed in real-world settings (for instance in schools [50, 103, 57, 18], homes [32] and care centres [35, 107]) 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 [35] or in the H2020 MuMMer [36] and FP7 Spencer [96] projects), or a narrow, well-specified complex task (for instance, supporting exercising in a gym, as I did in [107]). However, general purpose, long-term interaction is still an open question.

Novely, 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.

change 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: **This resaerch programme will lead to the design, implementation and real-world demonstration of the AI engine of a socially-intelligent robot. 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 deployed co-designed autonomous robots in several real-world, 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, which leads to a fourth ambition: establishing myself as one of the key leaders in social robotics. Surprisingly few groups worldwide have achieved full autonomy for a complex social robot – the LAAS is 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 [80, 82, 109] – 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 [82], 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 suffi-

ciently complex and open to explore novel social interactions. My objective is also to show that this real-world deployment can be successfully driven by the 'end-to-end' involvement of all the end-users and stakeholders: from defining the robot's role, from the different perspective of each end-user, to actually designing and 'teaching' the robot what to do.

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

Research strands

I present in this section the four main research strands that I intend to develop as a CNRS research scientist. These research priorities are guided by the scientific objectives listed in the previous section:

- **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 like 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.
- Finally, **Strand 4** will both: build-up the **experimental capacity** of the LAAS-CNRS in HRI with large scale field deployments; and simultaneously co-construct with the general public and ethics expert the conceptual framework and concrete guideline for **responsible human-robot interactions**.

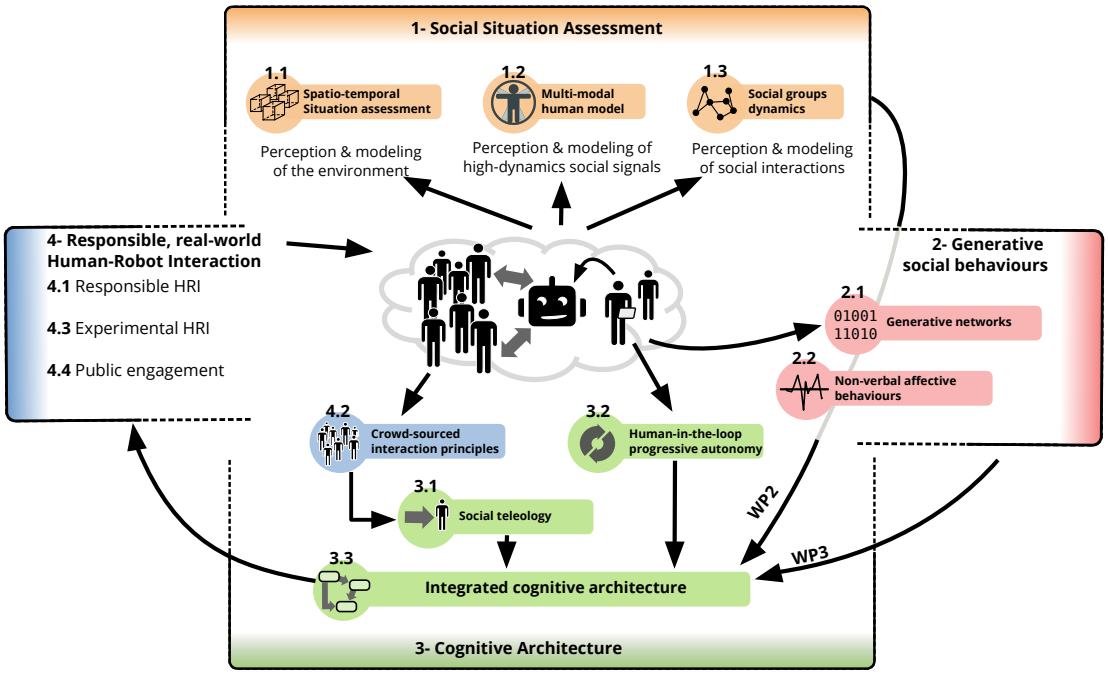


Figure 1.1: Overview of the research strands that I intend to develop as a CNRS research scientist.

These four strands are tightly coupled (Figure 1.1), and together, will enable a major scientific and technical breakthrough in autonomous, socially-intelligent, robots.

More specifically, Figure 1.1 gives an overview of the project workpackages, and their interrelations. Fieldwork plays a central role in the project, and appears in the centre of the figure. The first important field deployment is a one-year experiment, taking place at the Bristol science centre (T1.1). This 'public-in-the-loop' experiment is analysed and lead to the definition of core interaction principles (T1.2). These are in turn translated into algorithmic models, guiding the social teleology of the cognitive architecture (T4.1).

This first experiment is immediately followed by two other long-term experimental deployments: a one-year deployment in one of Bristol's Special Education Need (SEN) school (T5.1), followed by a one-year deployment at Bristol's Children's hospital (T5.2). These two additional experiments are both inputs for WP2 and WP3, and demonstrator for the robot socio-cognitive architecture (WP4).

Specifically, workpackage WP2 research, develop, and integrate all the components pertaining to the assessment of the spatio-temporal and social environment of the robot. Reference interaction situations and the data required to support this workpackage is directly drawn from the experimental fieldwork that will take place at the same time in WP1 and WP5. The perceptual capabilities delivered by WP2 are continuously integrated into the robot's cognitive architecture (T4.3), iteratively improving the socio-cognitive performances of the robot.

Workpackage WP3 looks into behaviour generation using machine learning (T3.2) and non-verbal affective modalities (T3.3). T3.2 is data-intensive, and will use datasets acquired during the field deployments (T1.1, T5.1, T5.2), as well as lab-recorded dataset of social interactions. Similar to WP2, the capabilities built in WP3 are integrated in the robot architecture in T4.3.

In addition to the integration of WP2 and WP3 capabilities, WP4 is also researching and developing the socio-cognitive drives of the architecture. They come both from T1.2 (as previously mentioned), and human-in-the-loop/public-in-the-loop machine learning (T4.2). T4.2, in particular, is tighly connected to the experimental fieldwork, where the learning-from-end-users take place.

Strand 1: Framing robot–supported human–human interaction

The basic ambition of **change** 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). WP1 operationalises this objective in two tasks: a theoretical contribution, examining the interplay between r-HHI, responsible AI, and ethics; and a large-scale study to gather public input.

T1.1 – Conceptual framing of r–HHI and ethical framework The first task in WP1 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?

Each of the field experiments (T1.2, T5.1. T5.2) will both *build on* and *feed into* the framework developed in this task. The work of this task will be structured around four two-days workshops, spread over the duration of the project (see Gantt chart). During these workshops, the **change** Ethics Advisory Board, local ethics experts (including the head of

the university ethics committee), and the **change** experimental partners (WeTheCurious, the SEN school network, the Children's hospital) will meet to debate and iterate over ethics guidelines for responsible long-term social interactions with robots.

Main outcomes of T1.1: a conceptual framework that clarify and organise together the questions raised by long-term social interactions; initial ethical guidelines for such interactions, aimed at informing future policy making.

T1.2 – 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, T1.2 will create and deploy a novel investigation methodology that I term 'experimental crowd-sourcing'. For one year, in close partnership with the Bristol Science centre WeTheCurious and its 'City Lab' programme, the visitors of the science centre will be invited to teleoperate a **change** robot, 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[82] and[109]), 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 at the science centre to reflect and iterate on these principles.

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.

Main outcomes of T1.2: a set of crowd-sourced interaction patterns and principles, that will inform the long-term social goals of the robot (T4.1); a large dataset of social interactions to feed into WP2 and WP3.

Technical work packages: WP2, WP3, WP4

The technical work programme of **change** is spread over work packages WP2, WP3 and WP4. Figure 1.2 gives an overview of the whole AI engine. WP2 (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. WP3 (bottom) significantly improve upon techniques for non-repetitive, socially-congruent behaviour production, combining recent advances in generative neural nets, art, and novel acoustic communication modalities. WP4 (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.

Strand 2: Real-world Social Situation Assessment

WP2 will integrate a full representation system for the social environment of the robot. It builds on existing state of art in *situation assessment* and *knowledge representation* (T2.1), and extend it to the social sphere (T2.2, T2.3 and T2.4).

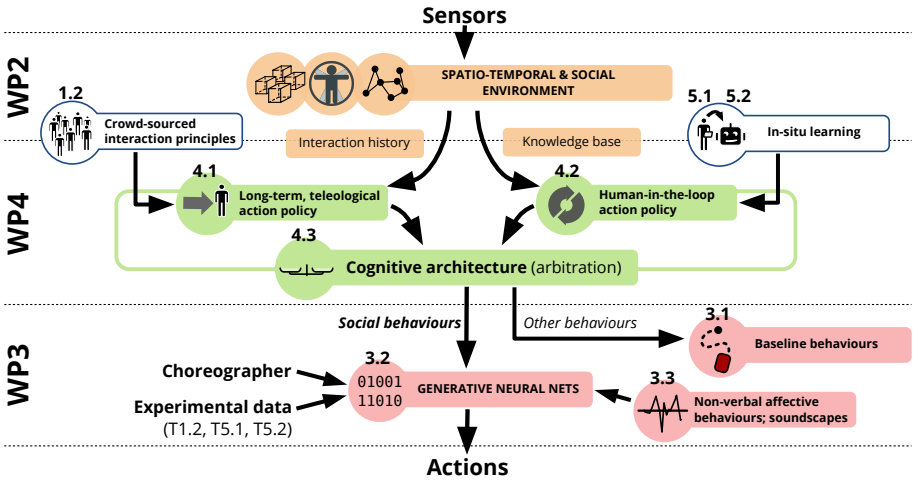


Figure 1.2: Overview of the AI engine implemented in **change**.

T2.1 – Hybrid situation assessment and knowledge representation Knowledge representation and grounding is a fundamental building block for cognitive architectures[59, 9]. This task builds on existing work on symbolic knowledge representation (eg[93] or my own work[58]) and my work on situation assessment[64] (that includes for instance object recognition and physics simulation[79]), to create a coherent system of representations for the cognitive architecture that extends the underworlds spatio-temporal representation tool[64] with symbolic and hybrid (like *conceptors*[41]) representations capabilities.

Main outcomes of T2.1: an extensible multi-modal software platform, that tracks and represents the spatio-temporal environment of the robot (including the locations and objects in the robot vicinity).

T2.2 – Multi-modal human model This task focuses on the acquisition, processing and modelling of social signals[34] to build a multi-modal model of the humans in the robot's vicinity. I have recently introduced a dataset of social interaction[61] 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[6]. This dataset and the related methodologies on data-driven social modeling will form the basis of this task, with additional data of natural interactions collected during T1.2.

Main outcome of T2.2: A data-driven social signal processing pipeline to model the surrounding humans.

T2.3 – Interaction and group dynamics Building on T2.2, T2.3 investigates the automatic understanding and modelling of group-level social interactions[92], including f-formations[68], sociograms (as done in[30] for instance), and inter-personal affordances[74]. This task builds on literature on on social dynamics analysis (eg[24, 42, 70]) to apply it to real-time social assessment by a robot, itself embedded into the interaction.

Main outcome of T2.3: the software pipeline required for the automatic analysis of social dynamics at group-level, able to model in real-time the social context of the robot.

T2.4 – Social situation assessment In T2.4, I integrate the social cues from T2.2 and T2.3 into the representation platform of T2.1. 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 representation capabilities of T2.1 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[26] (as I previously framed in[53, 22]).

Part of this task, I will also construct a *social embedding* of the robot: a compact, low-dimensional representation of the full social environment, that can be easily integrated with the machine learning algorithms developed in WP3 and WP4.

A focused experimental programme accompanies T2.4, to demonstrate (in relative isolation) the resulting socio-cognitive capabilities. I will implement a subset of the experimental protocols identified by Frith and Happé[29] to investigate theory of mind with autistic children, as it offers an excellent experimental framework for social robotics[53] for this work.

Main outcome of T2.4: a novel cognitive sub-system for social situation assessment, released as an open-source set of integrated ROS modules. This tool 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.

Strand 3: Generative social behaviours

Mirroring WP2's focus on understanding the social interactions, WP3 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).

Using on-board speech recognition (Mozilla DeepSpeech), the robots will be able to understand and record the textual transcription of the what the end-users say (in WP5, mostly children). The robots themselves are however purposefully designed *not* to speak, using instead non-verbal communication mechanisms (non-verbal utterances using sounds, gaze, joint attention, expressive motions, etc). This is a critical interaction design choice, that ensures we can more effectively manage what cognitive capabilities are ascribed to the robot by the users (expectation management). **change** seeks 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.

T3.1 – Behavioural baseline T3.1 establishes a baseline for behaviour generation, by surveying and implementing the current state of the art. In addition to traditional approaches like behaviour libraries, this will cover techniques like curiosity-driven behaviours[72], Learning from Demonstration[11, 3], human-in-the-loop action policy learning[84, 82]. This baseline will enable early in-situ experimental deployments (WP5), while also provide a comparison point for T3.2.

Using activity switching to support long term engagement with diabetic children[18]

Main outcomes of T3.1: A set of base behaviours for the robot, both social (like gesture, gaze), and generic (like navigation in crowded space). This task focuses on providing a working set of robot behaviours early in the project, using existing state of art.

T3.2 – Generative neural network for social behaviour production

Producing non-repetitive social behaviours is an open research question. I aim at significantly advancing the state of the art in this regard, by combining two recent techniques: (1) generative neural networks for affective robot motion generation[**yang2019appgan**, 67, 90]; (2) interactive machine learning in high-dimensional input/output spaces, where I have shown with my students promising results for generating complex social behaviours[82, 109] that fully involve the end-users[110].

In[90], 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) with a dataset co-created with a choreographer: during one month, a choreographer from the puppetering company RustySquid (with whom we have had several collaborations) will join the lab and remotely 'puppet' the robot while interacting with the lab members. The aim will be to collect a large amount of data to train the 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 T2.4: the generated behaviours will be shaped by the current social state of the interaction.

Main outcomes of T3.2: 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.

T3.3: Non-verbal behaviours and robot soundscape

In task T3.3, we introduce a novel non-verbal interaction modality for robots, based on soundscapes: 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[33]). The soundscapes that we will 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.

Main outcomes of T3.3: the development and implementation of soundscapes, a novel non-verbal interaction modality, integrated with the behaviours production of T3.2.

Strand 4: Goal-driven socio-cognitive architecture

WP4 design and implement on the R1 robot the principled cognitive architecture that binds together the socio-cognitive perceptual capabilities of the robot (WP2), with its action production mechanisms (WP3).

T4.1 – A social teleology for robots *Teleological systems* (ie goal-driven) has 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 [**moulinfrier2014self**, 72]or virtual agents[**pathak2017curiosity**]. This task's objective is to define and implement a novel *social teleology* that would algorithmically encode long-term social goals into the robot. This will directly build from the results of WP1, where interaction principles for social robots are experimentally uncovered.

Main outcomes of T4.1: the algorithmic translation of WP1's interaction principles in long-term social goals for the robot, eg a long-term, socially-driven action policy for the robot.

T4.2 – Learning from humans to achieve 'by-design' responsible & trustworthy AI Building on my recent, promising results on human-in-the-loop social learning[82, 109], this task implements the learning mechanics (including the bi-directional interface between the human teacher and the robot) to allow human end-users to teach the robot domain-specific (at school, at the hospital) social policies, following the methodology and the interactive reinforcement learning approach I developed with my students in[80].

In addition, this task will study through qualitative methods (thematic interviews and questionnaires) 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.

Main outcomes of T4.2: a human-in-the-loop reinforcement learning paradigm, suitable for in-situ teaching of the robot by the end-users themselves.

T4.3 – Integrating a socially-driven architecture for long-term interaction This task builds on the state of art in cognitive architectures (disembodied ones[17, 98, 45, 23, 49, 91, 94], as well as ones specifically developed for robotics: ACT-R/E[95], HAMMER[20], PEIS Ecology[77, 19], CRAM/KnowRob[9, 93], KeJia[16], POETICON++[2], and my own, the LAAS Architecture for Social Interaction[59]): the overall purpose of the socio-cognitive architecture of **change** is to integrate in a principled way the spatio-temporal and social knowledge of the robot (WP2) with a decision-making mechanism, to eventually produce socially-suitable actions (WP3).

The decision-making mechanism is the heart of the **change** AI engine: 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 in T1.2, I will combine long-term, socially-driven goals (*social teleology*, T4.1), and human-in-the-loop machine learning (T4.2) using a novel arbitration mechanism.

to make ensure local adaptation progressively learn an social policy enabling long-term autonomy. This task focuses on 'bringing the pieces together' in a principled manner.

The arbitration mechanism itself will build on research on reinforcement learning for experience transfer[65] that enables the re-assessement of a policy (here, our long-term social teleology) based on specific experience (here, the end-user-taught policy).

Main outcomes of T4.3: A cognitive architecture, implemented on the R1 robot, that enables long-term social engagement, by combining long-term goals with domain-specific action policies, taught by the end-users themselves.

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

change has the ambition to demonstrate long-term, co-designed social interactions in two complex, socially sensitive spaces. The first one involves the deployment of social robots in

special needs schools (SEN schools) in Bristol (T5.1). Building on a rigorous participatory approach involving the school teachers, as well as the parents, we will seek to integrate the robot in the daily life of the school, supporting the development of the students' physical and social skills. The second one takes place in Bristol's Children's Hospital (T5.2), supporting isolated children who suffer long-term conditions, in close cooperation with the hospital staff. In both cases, a social robot will be deployed on premises, for one uninterrupted year. It will integrate the daily routines of the institutions, under supervised autonomy[80], and *without* requiring the presence of a researcher at all time.

These two experiments raise specific practical and ethical questions, as they target vulnerable populations. This is an however informed choice: first, I already have established partnerships with Bristol's children hospital on one hand, and a network of Bristol-based SEN schools on the other hand. As such, and from a practical perspective, I do not foresee any institutional issues – on the contrary, our partners are excited at the prospect of taking part to the project. Besides, convincingly demonstrating the importance and positive impact of socially-driven, socially-responsible robotics does accordingly require complex social situations, and complex social dynamics. The two scenarios, which complement each other, provide both. These scenarios also put the project in the unique position of actually delivering high societal impact: we anticipate 30+ hospitalised children with long-term conditions, and 250+ SEN-educated children to directly benefit of the project, showing how robots can have a lasting, beneficial impact on the society, alongside human carers: it will establish the idea of *robots supporting human interactions* instead of dehumanising our social relationships.

Both these deployments will take place within the strict ethical framework established in T1.1, the ethical considerations pertaining to these experiments are further discussed below, in the section on ethics, and in the separate annex on ethics, uploaded alongside this proposal.

explain that these 2 large experiments will be scaffolded by many smaller ones

T5.1: A robot companion to support physical, mental and social well-being in SEN schools

Inspired by a similar large-scale deployment of social robots in Hong-Kong's SEN schools[75], the first study investigate whether a socially assistive robot can effectively support the development, social interactions and well-being of children with a long-term mental condition. This study will take place within the network of Bristol-based SEN schools, with which I already have an on-going collaboration. Specifically, the two main questions we seek to investigate are: What are the social underpinnings of the successful integration of a social robot in the school ecosystem? Can ambitious co-design with the end-users (teachers) deliver a 'net gain' for the learning, social interaction and well-being of the students?

The core of the study consists in deploying the R1 social robot in one of Bristol-based SEN school (Mendip Primary School, with possible extensions to other schools), to investigate how the robot can help shaping a social school ecology that fosters mental well-being, while effectively supporting teachers and students in their learning.

The study will adopt a strong participatory design approach, inspired by Patient and Public Involvement methodologies (PPI[13]), with 3 one-day focus groups organised with the school teachers; two evening focus group with the school parents, prior to the study; and several preparatory workshop at the school premises to involve the students as well.

The school study itself will take place during Y3, with the robot permanently based

at the school. The robot will take part in the regular teaching and other daily routines of the school, and will directly interact with the children, learning its action policy ('when to do what') from initial co-design with the teachers, followed by progressive in-situ teaching (see T4.2).

During selected 'observation days', observations will be conducted by the research team, and regular semi-structured interviews will be conducted with the teachers, parents, and where possible, the children themselves (using engagement metrics like the Inclusion of Other in Self task and Social-Relational Interviews[103]), to understand how the robot impacts the school dynamics (both positively and potentially negatively).

The task will be jointly supervised with local colleague and expert Dr.Nigel Newbutt, who has a long track record of working with special needs schools.

T5.2 – A robot companion to support isolated children during their hospital stay

The second experiment will take place within the paediatric ward for long-term conditions at the Bristol Children's Hospital. The ward has 8 beds, with children staying from a few weeks to several years. Over the course of the one-year deployment, we expect the robot to interact with about 30 children, their parents, and the hospital staff (nurses, doctors).

Similar to the first experiment, we will be using a *mutual shaping* approach[110] to design the role of the robot with the different stakeholders (nurses, doctors, parents, children), in order to experimentally investigate how a social robot can support hospitalised children with long-term conditions. The robot's role will revolve around facilitating social interactions between (possibly socially isolated) children, by fostering playful interaction within the paediatric ward.

This second experiment complements the first one by evidencing the commonalities and divergences in terms of social interactions when the robot is moved to a different environment. While the hospital eco-system is comparatively smaller that the SEN school one, people 'live' at the ward day and night; it becomes *de facto* the second home of the children, and the children will have more interaction opportunities than at the SEN school (where the robot is shared amongst a larger group). As a consequence, we expect to observe different interaction patterns, with potentially deeper affective engagement between the robot and the other ward's 'inhabitants'. Specific safeguarding measures will be put in place with the hospital team, and resulting observations will feed into the ethical guidelines of T1.1.

Experimental approach

My experimental approach has two phases. First, I will co-design and co-construct the robot's social role and behaviours through large-scale public engagement. For a whole year, I will deploy the **change** robot within the Open City Lab of Bristol Science Centre *We The Curious*, relinquishing its control to the visitors themselves. Tasked with remotely operating the robot to assist fellow visitors, a researcher will accompany them in 'inventing by doing' a new grammar of social interactions to develop answers to the questions: what does it mean for a robot to help? How to do so in the dynamic, messy, environment of a science centre? What are acceptable behaviours? Can we see new social norms emerge? At the end of this experiment, we expect 1000s of people to have experienced – and co-designed – how robots should interact with humans in a positive, helpful way. Each of these experiences will contribute to uncovering and designing the basic principles of social interaction for robots. This work is the focus of WP1.

While most of the interactions in the science centre will be short-lived, a follow-on, long term (one year) experiment will take place in one of Bristol's Special Education Needs (SEN) school where I currently run pilots, helping 250+ children with psycho-social impairments (autism) to develop their social skills and to engage into playful social activities: telling stories, triggering group activities with other children, providing additional social presence. Similar to the science museum experiment, the robot behaviours will be co-designed with, and learnt from the end-users themselves: teachers, parents, and as much as possible, the children themselves.

Importantly, **change** focuses specifically on the AI engine of the robot: I will use an existing robotic platform (Halodi's EVE, pictured on the left) and develop and train the algorithms required to achieve autonomy and responsible, long-term social utility. Indeed, after an initial training period, the robot will be autonomous: while the users will be provided with tools to override the robot's decisions at any time (via both an app and touch sensors on the robot itself), it will otherwise move and act on its own, without the need for constant supervision.

Ethics considerations and measures to ensure Responsible Research and Innovation

The **change** 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 **change** studies 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?).

Background on social robotic ethics

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*[106])?
- 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[4] 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 **change** 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 **change** will 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[88, 15], 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[69, 86]. As an example, a common objection to human-robot interaction is the perceived deceptive nature of the robot's role. It has been argued[12] 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."[1]. Specific subfields, like assistive robotics[85], have seen some additional work, but social

robotics is still not equipped with operational guidelines, similar to the EU guidelines on trustworthy AI.

change–specific measures

I have chosen to focus the first work package task (T1.1) on building an operational ethical framework for social robots which engage over long period of time with the public. This work will deliver initial guidelines – strongly inspired by the guidelines on Trustworthy AI – that will both form the ethics basis for the **change** experimental fieldwork, and have an impact beyond the project, to feed into future European-level guidelines.

This work will be supported by an Ethics Advisory Board, composed of 3 experts in ethics and social robotics and AI. While the exact composition of the board is not final yet, it will include at least one member from the EU High-Level Expert Group on Artificial Intelligence, that will be able to share the EU expertise in framing ethics guidelines.

Practically speaking, these guidelines will form the basis of the ethics approval process for the three long-term **change** studies. It will be additionally supported by my extensive experience in seeking ethics approval for studies involving robots and vulnerable populations (in particular, children[57, 61, 82]), the expertise of Dr. Newbutt in conducting research with SEN schools (T5.1), and the support of J. Bowyer at the Bristol's Children's Hospital to obtain NHS ethics approval. **As per requested, details of the ethics approval process, children safeguarding, research Code of Conduct, and Data Management Plan are annexed to the project proposal, in a separate 'Ethics and Data Protection' document**.

The project will also follow the European Commission recommendations for Responsible Research and Innovation (RRI). RRI is defined in[89] (and has been subsequently adopted by the UK Engineering and Physical Sciences Research Council[73]) using the acronym AREA: Anticipation, Reflection, Engagement and Action. The **change** research will be undertaken responsibly by (1) Anticipating possible consequences; (2) by integrating mechanisms of Reflection about the conducted work and its aims; (3) by Engaging with relevant stakeholders (general public, teachers, hospital staff, parents, children themselves); and (4) by Guiding action of researchers accordingly. This approach has been formalised in the AREA 4P framework[87] [1], that I will use to guide the research strategy over the course of the project. An additional role of the Ethics Advisory Board will be to advise and audit the project with regards to this framework for responsible research.

Risk/gain assessment; risk mitigations

Tasks 1.1, 1.2 develop a novel methodology, 'public-in-the-loop' machine learning, for large-scale co-design of social interactions with the public. If successful, this will be of great value, well beyond the project. The proposed experimental setup (science centre visitors 'taking control' of the robot) might however lead to interactions that are either too short or to artificial to create meaningful, generalisable social interaction. In addition, the messy and complex nature of the science centre environment is also currently beyond-state-of-the-art in term of extracting the useful social features required to train a classifier.

However, the interaction principles that we want to uncover in T1.1 and T1.2 (and that are feeding into WP2 and WP4) will principally come from a qualitative analysis of the interactions, carried in parallel to the machine learning approach. This well within the

[1] https://www.orbit-rri.org/about/area-4p-framework/

expertise of the PI, and, as such, is low-risk. T1.1 can thus be described as a **medium–risk, high–gain** component of **change**.

Task 2.1 develops a novel situation assessment component, that integrates spatio-temporal modeling with knowledge representation. The resulting component is beyond-state-of-the-art, and would be highly relevant to a large range of robotic applications. This component relies on integrating tools that are independently relatively mature and well understood, and the principles of the integration itself is already well researched. Besides, it falls well within the PI expertise [64, 79, 58]. As such, T2.1 can be described as **low–risk, medium–gain**.

Tasks 2.2, 2.3, 2.4 Work on real-time modeling of social dynamics in real-world environments are only begining to be studied in robotics. While the underpinning are well understood in neighbouring academic fields, a very significant work remain to be done to integrate disparate or partial approaches into one framework. These tasks also require the acquisition of novel datasets that focus on natural human-human social interactions. The PI has extensive experience in building and acquiring such datasets [61, 78], and does not foreseen major difficulties. The resulting components have however the potential to unlock a new class of social robots, aware in real-time of their social surroundings and dynamics. These tasks are thus considered **low–risk, high–gain**.

Task 3.1 The behavioural baseline implements the current state-of-the-art, and as such is **low–risk, low–gain**. T3.1 will guarantee early on in the project a 'working' robot, yet with predictable/repetitive behaviours.

Task 3.2 The neural generation of complex social behaviours is a **medium–risk, high–gain** task: while it builds on solid existing state-of-the-art, it relies on very significant progress in both the modeling of the social dynamics (WP2) and the capacity of designing a machine learning approach to learn and generate these complex behaviours. While the former falls well within the PI expertise, machine learning for social motion generation is essentially a novel field. The success of this task will rely to a large extend on the quality of the post-doctoral researcher recruited to lead this effort. The main mitigation to the risk associated to T3.2 is the behavioural baseline created in T3.1: the behavioural capabilities generated in T3.2 can be complemented by ad-hoc behaviours whenever required.

Task 3.3 Non-verbal communication is a well established subfield of HRI research, well known to the PI. The creation of the novel interaction modality based on soundscape is novel, with potential for impact beyond the project. This new modality will be co-developped with an expert of sound design for interaction, and we do not foresee major risks. Overall, the task is **low–risk, medium–gain**.

Task 4.1 The conceptual framing of a *socially–driven architecture* (social teleology) and its translation into decision-making algorithms are to a large extend open questions. This task might however lead to uncover a fundamental mechanism to enable long-term engagement of users with social robots. Building fundamentally on blue-sky research, this task is **high–risk, high–gain**. If not successful, I will instead rely on the decision-making strategy of T4.2, which is much lower risk.

Task 4.2 The techniques developed in T4.2 have been previously used and tested by the PI in two different real-world environments [82, 107]. While they will require significant adjustments for this project, the task is overall **low–risk, low–gain**.

Task 4.3 The integration of the different cognitive functions of the robot into one principled cognitive architecture, that include cognitive redundancy, is one of the core expertise of the PI [59]. This task however includes significant novel elements (cognitive mechanisms for long-term autonomy; decision arbitration) that bear unknowns. Besides, this task is a critical pre-requisite for WP5. As a result, T4.3 is considered as **high-risk**. The task is focused on integration to meet the requirements of the WP5 experiments, and parts of the resulting software architecture might be project-specific. However the overall aims of endowing the robot with long-term social autonomy would be a significant breakthrough, and as such, T4.3 is **high-gain**. The main mitigations comes from (1) the iterative development process of the architecture, that will start from the existing state-of-the-art, to which the PI has previously contributed [59]. By doing so, a decisional architecture for the robot will be available early on in the project. While that architecture might be a scaled-down version of the initial ambition, it will still enable the fieldwork proposed in WP5, possibly with a lesser level of autonomy; (2) the possibility of using only one of the two action policies (T4.1 or T4.2), thus removing the need for complex arbitration.

WP5: Experimental deployments

The two application scenarios (at the children hospital and in the SEN school) are ambitious and inherently risky, as they target vulnerable populations. However, first, demonstrating the importance of advanced social modelling, and convincingly proving the effectiveness of our approach does require accordingly complex social situations, and complex social dynamics. The two scenarios, which complement each other, provide both.

Second, working with vulnerable populations, in constrained and complex environments (children hospital and SEN schools) adds significant risks to the project. But it is also what make the project in the unique position of delivering a high societal impact: a direct positive impact on children's lives (we anticipate 100+ hospitalised children and 50+ children with psycho-social impairements interacting over long periods of time with a robot over the course of the project), and a broader impact on the society, showing how robots can have a lasting, strong, positive impact on the society, also establishing the idea of *robots supporting human interactions* instead of dehumanising our social relationships.

Together, Task 5.1 and 5.2 are high-risk, high-gain.

The two main mitigations are (1) early and continuous engagement with the stakeholders, and (2) the decoupling of the two applications, meaning that the risks associated to each of them do not impact the other one.

Early engagement will be ensured by relying on a participatory design methodology, involving all the stakeholders from the onset of the project; the methodology will involve regular joint workshops; on-site (hospital and SEN schools) research stay including engagement with the staff/charities and the children themselves; early field testing and prototyping, relying if necessary on provisional, yet well-known, robot platforms available at the host institution (for instance, Softbank Nao and Pepper). This user-centered approach will be championed by the post-doc recruited on the project on WP4 and WP5, who will have to have a strong expertise in user-centered design.

It is also important to note that, while preparing this bid, initial discussions have been held with all the partners involved with the experimental fieldwork (WeTheCurious science centre, Bristol's Children Hospital, the network of SEN schools): each of these institutions is enthusiastic about the project, already contributing ideas to integrate the robots in their

daily routines, and ready to dedicate time and effort for its success.

Bibliography

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

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

[3] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. "A Survey of Robot Learning From Demonstration". In: *Robotics and Autonomous Systems* 57.5 (2009), pp. 469–483.

[4] 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.

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

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

[7] P. Baxter, E. Ashurst, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. "The Wider Supportive Role of Social Robots in the Classroom for Teachers". In: *WONDER Workshop, 2015 International Conference on Social Robotics*. 2015.

[8] P. Baxter, S. Lemaignan, and G. Trafton. "Workshop on Cognitive Architectures for Social Human-Robot Interaction". In: *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*. 2016. doi: 10.1109/HRI.2016.7451865.

[9] 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.

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

[11] A. Billard, S. Calinon, R. Dillmann, and S. Schaal. "Robot Programming by Demonstration". In: *Springer Handbook of Robotics*. Springer, 2008, pp. 1371–1394.

[12] 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.

[13] A. Boivin, K. Currie, B. Fervers, J. Gracia, M. James, C. Marshall, C. Sakala, S. Sanger, J. Strid, V. Thomas, et al. "Patient and public involvement in clinical guidelines: international experiences and future perspectives". In: *Quality and Safety in Health Care* 19.5 (2010), e22–e22.

[14] 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).

[15] 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.

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

[17] 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.

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

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

[20] 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.

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

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

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

[24] 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.

[25] J. Fink, P. Rétornaz, F. Vaussard, F. Wille, K. Franinović, A. Berthoud, S. Lemaignan, P. Dillenbourg, and F. Mondada. "Which Robot Behavior Can Motivate Children to Tidy up Their Toys? Design and Evaluation of "Ranger"". In: *Proceedings of the 2014 Human-Robot Interaction Conference*. 2014.

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

[27] R. Flook, A. Shrinah, L. Wijnen, K. Eder, C. Melhuish, and S. Lemaignan. "On the Impact of Different Types of Errors on Trust in Human-Robot Interaction: Are laboratory-based HRI experiments trustworthy?" In: *Interaction Studies* (2019). doi: 10.1075/is.18067.flo.

[28] S. Forestier and P.-Y. Oudeyer. "A Unified Model of Speech and Tool Use Early Development". In: *39th Annual Conference of the Cognitive Science Society (CogSci 2017)*. Proceedings of the 39th Annual Conference of the Cognitive Science Society. London, United Kingdom, July 2017. url: https://hal.archives-ouvertes.fr/hal-01583301.

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

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

[31] M. Gharbi, S. Lemaignan, J. Mainprice, and R. Alami. "Natural Interaction for Object Hand-Over". In: *Proceedings of the 2013 ACM/IEEE Human-Robot Interaction Conference*. 2013.

[32] 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.

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

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

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

[36] 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.

[37] 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.

[38] D. Hood, S. Lemaignan, and P. Dillenbourg. "When Children Teach a Robot to Write: An Autonomous Teachable Humanoid Which Uses Simulated Handwriting". In: *Proceedings of the 2015 ACM/IEEE Human-Robot Interaction Conference*. 2015.

[39] B. Irfan, J. Kennedy, S. Lemaignan, F. Papadopoulos, E. Senft, and T. Belpaeme. "Social psychology and Human-Robot Interaction: an Uneasy Marriage". In: *Proceedings of the 2018 ACM/IEEE Human-Robot Interaction Conference*. 2018. doi: 10.1145/3173386.3173389.

[40] A. Jacq, S. Lemaignan, F. Garcia, P. Dillenbourg, and A. Paiva. "Building Successful Long Child-Robot Interactions in a Learning Context". In: *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*. 2016. doi: 10.1109/HRI.2016.7451758.

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

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

[43] J. Kennedy, S. Lemaignan, and T. Belpaeme. "The Cautious Attitude of Teachers Towards Social Robots in Schools". In: *Proceedings of the 21th IEEE International Symposium in Robot and Human Interactive Communication, Workshop on Robots for Learning*. 2016.

[44] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft, and T. Belpaeme. "Child Speech Recognition in Human-Robot Interaction: Evaluations and Recommendations". In: *Proceedings of the 2017 ACM/IEEE Human-Robot Interaction Conference*. 2017. doi: 10.1145/2909824.3020229.

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

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

[47] S. Lallée, U. Pattacini, J. Boucher, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. Sisbot, G. Metta, R. Alami, M. Warnier, J. Guitton, F. Warneken, and P. Dominey. "Towards a Platform-Independent Cooperative Human-Robot Interaction System: II. Perception, Execution and Imitation of Goal Directed Actions". In: *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2011.

[48] S. Lallée, U. Pattacini, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. A. Sisbot, G. Metta, T. Pipe, R. Alami, M. Warnier, J. Guitton, F. Warneken, and P. F. Dominey. "Towards a Platform-Independent Cooperative Human Robot Interaction System: III. An Architecture for Learning and Executing Actions and Shared Plans". In: *IEEE Transactions on Autonomous Mental Development* (2012).

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

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

[51] 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.

[52] S. Lemaignan and R. Alami. "A Few AI Challenges Raised while Developing an Architecture for Human-Robot Cooperative Task Achievement". In: *Proceedings of the AAAI 2014 Fall Symposium Series – Artificial Intelligence and Human-Robot Interaction*. 2014.

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

[54] S. Lemaignan, J. Fink, and P. Dillenbourg. "The Dynamics of Anthropomorphism in Robotics". In: *Proceedings of the 2014 ACM/IEEE Human-Robot Interaction Conference*. 2014.

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

[56] S. Lemaignan, F. Garcia, A. Jacq, and P. Dillenbourg. "From Real-time Attention Assessment to "With-me-ness" in Human-Robot Interaction". In: *Proceedings of the 2016 ACM/IEEE Human-Robot Interaction Conference*. 2016. doi: 10.1109/HRI.2016.7451747.

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

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

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

[60] S. Lemaignan and R. Alami. "Explicit Knowledge and the Deliberative Layer: Lessons Learned". In: *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2013.

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

[62] S. Lemaignan, J. Fink, F. Mondada, and P. Dillenbourg. "You're Doing It Wrong! Studying Unexpected Behaviors in Child-Robot Interaction". In: *Proceedings of the 2015 International Conference on Social Robotics*. 2015.

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

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

[65] 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.

[66] A. Mallet, C. Pasteur, M. Herrb, S. Lemaignan, and F. Ingrand. "GenoM3: Building middleware-independent robotic components". In: *Proceedings of the 2010 IEEE International Conference on Robotics and Automation*. 2010.

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

[68] 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.

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

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

[71] F. Mondada, J. Fink, S. Lemaignan, D. Mansolino, F. Wille, and K. Franinović. "New Trends in Medical and Service Robots". In: vol. 38. Mechan. Machine Science. Appeared first as a paper at MESROB2014. Springer Publishing, 2015. Chap. Ranger, an Example of Integration of Robotics into the Home Ecosystem. isbn: 978-3-319-23831-9. doi: 10.1007/978-3-319-23832-6_15.

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

[73] R. Owen. "The UK Engineering and Physical Sciences Research Council's commitment to a framework for responsible innovation". In: *Journal of Responsible Innovation* 1.1 (2014), pp. 113–117. doi: 10.1080/23299460.2014.882065. eprint: https://doi.org/10.1080/23299460.2014.882065. url: https://doi.org/10.1080/23299460.2014.882065.

[74] 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.

[75] *robot4SEN website*. 2019. url: https://translate.google.com/translate?sl=auto%5C&tl=en%5C&u=http%3A%2F%2Fwww.robot4sen.org%2F.

[76] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken. "Which One? Grounding the Referent Based on Efficient Human-Robot Interaction". In: *19th IEEE International Symposium in Robot and Human Interactive Communication*. 2010.

[77] 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.

[78] Y. Sallami, K. Winkle, N. Webb, S. Lemaignan, and R. Alami. "The Unexpected Daily Situations (UDS) Dataset: A New Benchmark for Socially-Aware Assistive Robots". In: *Companion Proceedings of the 2020 ACM/IEEE Human-Robot Interaction Conference*. 2020. doi: 10.1145/3371382.3378270.

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

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

[81] E. Senft, S. Lemaignan, M. Bartlett, P. Baxter, and T. Belpaeme. "Robots in the classroom: Learning to be a Good Tutor". In: *Proceedings of the 2018 HRI workshop R4L 'Robots for Learning'*. 2018.

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

[83] E. Senft, S. Lemaignan, P. Baxter, and T. Belpaeme. "Leveraging Human Inputs in Interactive Machine Learning for Human Robot Interaction". In: *Proceedings of the 2017 ACM/IEEE Human-Robot Interaction Conference*. 2017. doi: 10.1145/3029798.3038385.

[84] 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.

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

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

[87] B. C. Stahl. "Implementing Responsible Research and Innovation for Care Robots through BS 8611". In: *Pflegeroboter*. Ed. by O. Bendel. Wiesbaden: Springer Fachmedien Wiesbaden, 2018, pp. 181–194. isbn: 978-3-658-22698-5. doi: 10.1007/978-3-658-22698-5_10. url: https://doi.org/10.1007/978-3-658-22698-5_10.

[88] 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.

[89] J. Stilgoe, R. Owen, and P. Macnaghten. "Developing a framework for responsible innovation". In: *Research Policy* 42.9 (2013), pp. 1568–1580. issn: 0048-7333. doi: https://doi.org/10.1016/j.respol.2013.05.008. url: http://www.sciencedirect.com/science/article/pii/S0048733313000930.

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

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

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

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

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

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

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

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

[98] 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.

[99] C. Wallbridge, S. Lemaignan, and T. Belpaeme. "Qualitative Review of Object Recognition Techniques for Tabletop Manipulation". In: *ACM Human–Agent Interaction Conference*. 2017.

[100] C. Wallbridge, S. Lemaignan, E. Senft, and T. Belpaeme. "Generating Spatial Referring Expressions in a Social Robot: Dynamic vs Non-Ambiguous". In: *Frontiers in AI and Robotics* (2019). doi: 10.3389/frobt.2019.00067.

[101] C. Wallbridge, S. Lemaignan, E. Senft, and T. Belpaeme. "Towards Generating Spatial Referring Expressions in a Social Robot: Dynamic vs Non-Ambiguous". In: *Proceedings of the 2019 ACM/IEEE Human–Robot Interaction Conference*. 2019. doi: 10.1109/HRI.2019.8673285.

[102] M. Warnier, J. Guitton, S. Lemaignan, and R. Alami. "When the Robot Puts Itself in Your Shoes. Managing and Exploiting Human and Robot Beliefs". In: *Proceedings of the 21th IEEE International Symposium in Robot and Human Interactive Communication*. 2012.

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

[104] L. Wijnen, P. Bremner, S. Lemaignan, and M. Giuliani. "Performing Human-Robot Interaction User Studies in Virtual Reality". In: *Proceedings of the 2020 RoMAN Conference*. 2020. doi: 10.1109/RO-MAN47096.2020.9223521.

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

[106] 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.

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

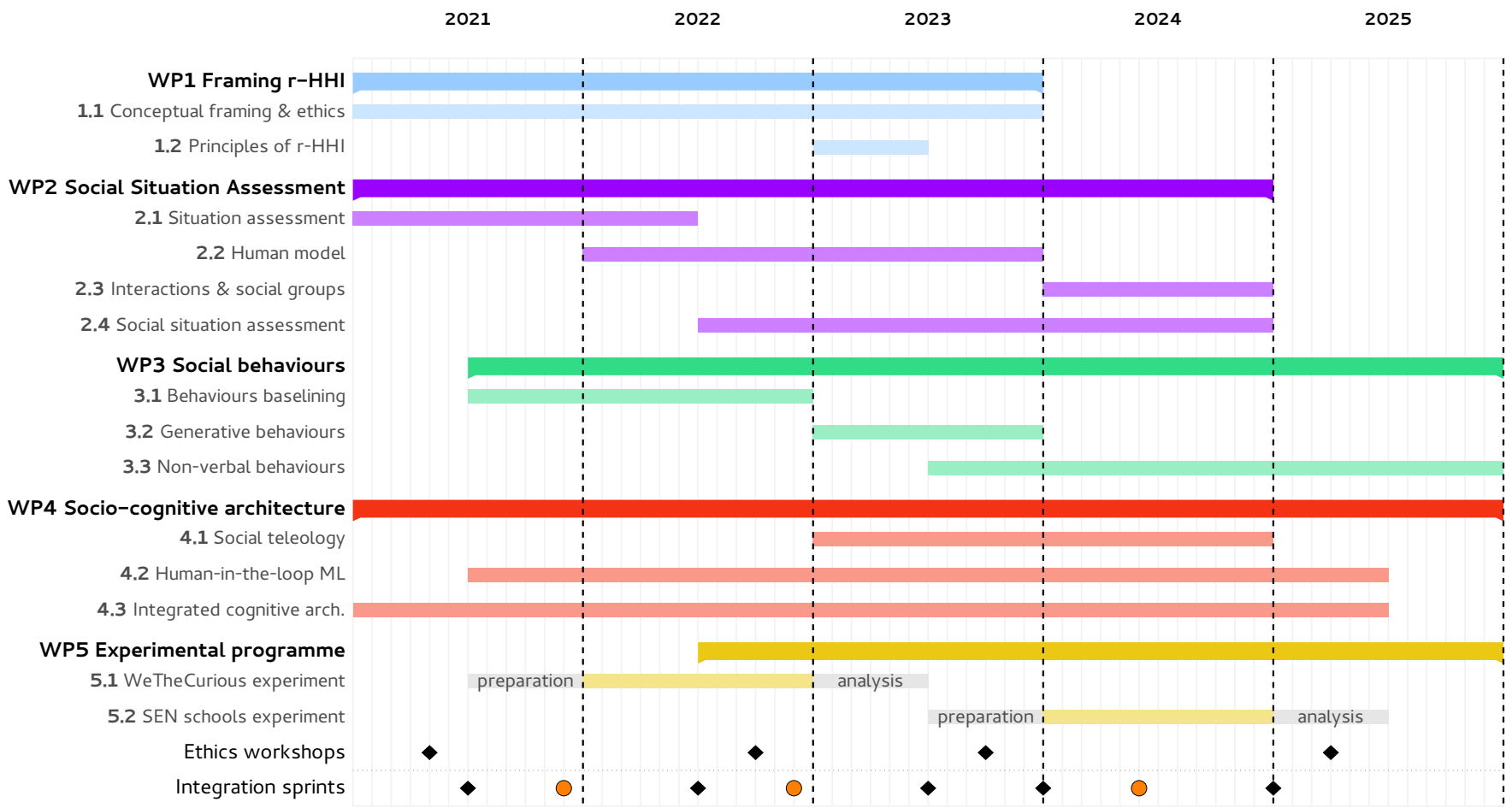
[108] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton, and P. Bremner. "Effective Persuasion Strategies for Socially Assistive Robots". In: *Proceedings of the 2019 ACM/IEEE Human–Robot Interaction Conference*. 2019. doi: 10.1109/HRI.2019.8673313.

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

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

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

Research plan for the first five years


2021
2022
2023
2024
2025
WP1 Framing r-HHI
1.1 Conceptual framing & ethics
1.2 Principles of r-HHI
WP2 Social Situation Assessment
2.1 Situation assessment
2.2 Human model
2.3 Interactions & social groups
2.4 Social situation assessment
WP3 Social behaviours
3.1 Behaviours baselining
3.2 Generative behaviours
3.3 Non-verbal behaviours
WP4 Socio-cognitive architecture
4.1 Social teleology
4.2 Human-in-the-loop ML
4.3 Integrated cognitive arch.
WP5 Experimental programme
5.1 WeTheCurious experiment
preparation
analysis
5.2 SEN schools experiment
preparation
analysis
Ethics workshops
Integration sprints

Importance and Integration in the scientific landscape

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

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

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

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 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

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

change delivers 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, the project builds on a **strong interdisciplinary team**: the post-docs directly recruited on **change** will have backgrounds in sociology of technology (PD1), cognitive modeling (PD2), machine learning (PD3), cognitive robotics (PD4). Additional expertise will be recruited to provide specific support: the **change** Ethics Advisory Board will contribute expertise to guide the work on ethics; Dr. Newbutt will provide expertise in learning technologies and cognitive impairment; Dr. Meckin will provide expertise in sound-based expressive communication; the WeTheCurious science centre will provide training in large-scale public engagement; the Bristol Children's hospital will bring the required expertise in working with young patients; the RustySquid company will provide expertise in expressive arts and puppetering.

Integration with the local research landscape

TODO, in particular connections to ANITI

- verbal behaviours -> collaboration within ANITI? - collaboration with Rafaelle

Academic track-record and contributions

In complement to the CV attached to my application, this section highlights my academic track-record and contributions to the generation of knowledge and development of individuals.

Academic profile

Since I completed my joint PhD in Cognitive Robotics from the CNRS/LAAS (France) and the Technical University of Munich (Germany), for which I received the *Best PhD in Robotics 2012* award from French CNRS and the prized *Cumma Summa Laude* distinction in Germany, I have emerged as a leading authority in HRI.

Soon after my PhD, I created and successfully led for 2 years the HRI group within the AI for Learning CHILI Lab at EPFL (Switzerland), supervising in total 10 students, and establishing in a short timeframe CHILI as an internationally recognised centre in robotics for education. While my original training was in **symbolic cognition & AI for autonomous robotics**, my postdoctoral stay at the highly cross-disciplinary CHILI Lab gave me the opportunity to become an expert in **experimental sciences, socio-psychology and education sciences**.

I was then awarded an EU **H2020 Marie Sklodovska-Curie Individual Fellowship** and I engaged in basic research on artificial cognition: over 2 years, I explored the underpinnings of artificial social cognition. I **contributed significantly to the framing of the emerging field of data-driven HRI**, also releasing of the PInSoRo open dataset (10.5281/zenodo.1043507), a **one-in-a-kind dataset of natural child-child and child-robot social interactions**.

My current role as a permanent **Associate Professor in Social Robotics and AI** at the Bristol Robotics Laboratory (BRL, largest co-located robotic lab in the UK) recognise my leadership. I am **in charge of defining and implementing the lab's research strategy in human-robot interactions**. I created the Embedded Cognition for Human-Robot Interactions (ECHOS) research group, that I now co-lead, supervising 15+ PhDs and post-docs. I also supervise the BRL's Connected Autonomous Vehicles research group (5 students and post-docs). Specifically, the ECHOS group covers most aspects of situated AI for human-robot interaction, **my role includes strategic planning of the group activities, scientific guidance, recruitment of staff and prospective students, and grant applications**.

My field of expertise covers **the socio-cognitive aspects of human-robot interaction, both from the perspective of the human cognition and the design and implementation of cognitive architectures for robots**. I have also focused a significant portion of my **experimental work on child-robot interactions in real-world educative settings**, exploring how robots can support teachers and therapists to develop engaging novel learning paradigms.

This expertise is recognised internationally: I have a substantial track record of aca-

demic outputs. Since 2008, I have authored or co-authored **75+ peer-reviewed publications** in international journals and conferences, leading to **2700+ citations**, h-index of 26, i10-index of 43 (source: Google Scholar).

I have established strong **peer recognition** in the field of human-robot interaction and cognitive robotics. For instance:

- invited to **high-profile editorial roles**: Programme Committee member of the HRI conference since 2015; editor of Frontiers In Robotics and AI journal; editor or Programme Committee member of several leading conferences in AI and Robotics (RSS, IROS, IJCAI, HAI, AAMAS);
- invited member of the UK EPSRC Peer Review College; invited reviewer for the French, Dutch, Israeli research agencies;
- numerous **invited talks** at national and international symposiums and events (9 invited talks since Jan. 2018, including **keynotes** at the UK Robotics and Autonomous Systems 2019 conference, and at the 2018 AAAI Fall Symposium);
- **local chair for the high-profile, international HRI2020 conference** (700+ delegates);
- regularly invited to PhD defense committees (most recently at LAAS, France, Uni Bielefeld, Germany, and Uni Örebro, Sweden).

I **actively engage with policy makers, at national and European level**: for instance, over the past 2 years, I have been directly interacting (through participating to panels, visits and one-to-one discussions) with the EU Research Executive Agency (MSCA AI Cluster 2019); the UK minister for Business, Energy and Industrial Strategy Greg Clark; the UK minister for Universities, Science, Research and Innovation Chris Skidmore; the chair of the West of England authority Tim Bowles; the UK Research & Innovation Portfolio manager for Robotics Clara Morri.

I have a **strong track record of tech transfer**, through patenting (US patent US20190016213A1) and involvement in national (UK) and EU-level projects focused on tech-transfer (InnovateUK ROBOPILOT, CAPRI, CAVForth; EU Terrinet, SABRE).

Finally, I actively engage in **research communication**: my past research has been covered several times by mainstream international media, including press releases by Reuters, Press Association; TV coverage by the BBC, Sky News; radio interviews and broadcast. My academic website (academia.skadge.org) showcases this media coverage. I also maintain an active, science-focused, presence on the social media (Twitter handle: @skadge).

Appropriateness of academic track record for the research programme

In addition to this academic profile, my publication track record puts me in a unique position to deliver on my proposed research programme. Table 3.1 lists, per domain, some of my academic outputs that are directly relevant to the research project.

I am also a technology expert, with major software and hardware contributions to the robotic community (including contributions to OpenCV and core components of Robot Operating System, ROS). As such, I have a clear grasp of the technical feasibility of the proposed work. I am also in the rare position of having substantial experience in designing and running full architectures for complex autonomous social robots [59, 109].

Table 3.1: PI's domains of expertise relevant to the research project

Psycho-social underpinnings of HRI	
human factors	anthropomorphism[54], cognitive correlates[55], social influence[108]
trust, engagement, social presence	[27][62][25][39][104]
theory of mind	perspective taking[76, 102], social mutual modelling[53, 22]
Social signal processing	
non-verbal behaviours	attention[56], child-child dataset[61], internal state decoding[6]
verbal interactions	speech recognition[44], dialogue grounding[63]
Behaviour generation	
social behaviours	[47], verbal interactions[100, 101], physical interactions[31]
interactive reinforcement learning	[83, 80, 82, 109]
Socio-cognitive architectures	
architecture design	[59, 8, 52, 48, 66]
knowledge representation	ontologies [58, 60]
spatio-temporal modelling	object detection [99], physics-aware situation assessment[64, 79]
Fieldwork in HRI	in classrooms [38, 57, 40, 7, 43, 81], at home [71], in public spaces [109]

Contributions to the generation of knowledge

Selected scientific outputs

Senft, E., Lemaignan, S., Baxter, P., Bartlett, M., Belpaeme, T.
Teaching robots social autonomy from in situ human guidance
Science Robotics 2019

A novel human-in-the-loop machine learning approach to implement social autonomy in a robot, with several deployments in UK public schools. This is a first-in-kind demonstration of learning autonomous action policy in a high dimensional, socially complex, environment.
[main study supervisor]

Wallbridge, C., Lemaignan, S., Senft, E., Belpaeme, T.
Generating Spatial Referring Expressions in a Social Robot: Dynamic vs Non-Ambiguous
Frontiers in AI and Robotics 2019

Challenges the common understanding that robots should be unambiguous: we show that ambiguity is often desirable for fluid and natural human-robot interactions.
[main study supervisor]

Bartlett, M., Edmunds, C. E. R., Belpaeme, T., Thill, S., Lemaignan, S. **What Can You See? Identifying Cues on Internal States from the Kinematics of Natural Social Interactions**
Frontiers in AI and Robotics 2019

Investigates how partially hidden 'internal states' (like emotions, cooperativeness, etc) can be decoded from simple visible cues, like skeletons. Also demonstrates that social situations can be described along 3 simple dimensions.
[main study supervisor]

Lemaignan, S., Edmunds E. R., C., Senft, E., Belpaeme, T.
The PInSoRo dataset: Supporting the data-driven study of child-robot social dynamics
PLOS ONE 2018

A first-in-kind, large scale dataset of child-child and child-robot social interactions. Design with machine learning in mind, this dataset effectively opens up the field of data-driven social psychology, with direct applications in AI and social robotics.**[principal investigator]**

Lemaignan, S., Sallami, Y., Wallbridge, C., Clodic, A., Alami, R.
underworlds: Cascading Situation Assessment for Robots
IEEE IROS 2018

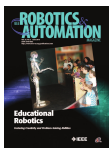
A novel representation technique to efficiently represent multiple parallel states of the world, including imaginary ones. This ability is critical to represent spatio-temporal predictions, and to create models of other agents' representations. **[principal investigator]**

Senft, E., Baxter, P., Kennedy, J., Lemaignan, S., Belpaeme, T.
Supervised Autonomy for Online Learning in Human-Robot Interaction
Pattern Recognition Letters 2017

The mathematical and technical bases of the SPARC paradigm for human-in-the-loop machine learning, showing that high-dimensional problems can be learnt effectively and rapidly thanks to an innovative input feature selection mechanism.
[student supervisor; 22 citations]

Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.
Artificial Cognition for Social Human-Robot Interaction: An Implementation
Artificial Intelligence 2017

Landmark article: one of the first complete, semantic-aware, robotic architecture for human-robot interaction, including symbolic knowledge representation, situation assessment, natural language grounding, task planning, human-aware motion planning and execution.
[principal investigator and coordinator; 143 citations]



Lemaignan, S., Jacq, A., Hood, D., Garcia, F., Paiva, A., Dillenbourg, P.
Learning by Teaching a Robot: The Case of Handwriting
Robotics and Automation Magazine 2016

Long-term studies with children and therapists, where we *reverse* the social role of the robot to significantly improve the children' self-confidence. A landmark in social robotics for education.
[principal investigator; 141 citations (incl. conf. article) **]**

Lemaignan, S., Ros, R., Sisbot, E. A., Alami, R., Beetz M. **Grounding the Interaction: Anchoring Situated Discourse in Everyday Human-Robot Interaction**
Intl Journal of Social Robotics 2012

In this paper, I show how symbolic knowledge representation can be used by robot to ground natural language interactions, also taking into account the unique perspective of the human interactor.
[principal investigator; 100 citations]

Lemaignan, S., Ros, R., Mösenlechner, L., Alami, R., Beetz, M.
ORO, a Knowledge Management Module for Cognitive Architectures in Robotics
IEEE IROS 2010

One of the very first knowledge base designed and integrated in service robots. Pioneering work which played a key role in understanding how intelligent robot can represent their knowledge to facilitate communication with humans.
[principal investigator; 158 citations]

Fellowships and awards

2019	UWE Vice Chancellor Accelerator Fellowship
2015 – 2017	**EU Marie Skłodowska-Curie Individual Fellowship** Theory of Mind and social robotics, Plymouth University, UK
HRI'2017	Best Paper award
HRI'2016	Best Paper award
AAAI'2015	Best Video award in Artificial Intelligence
HRI'2014	Best Late Breaking Report award
2012	**Best PhD in Robotics 2012** award, CNRS, France
2012	PhD with **High Distinction** ("Summa Cum Laude"), TU Munich
Ro-Man'2010	Best paper award

Contributions to the development of individuals

Since 2018, I co-supervise the Bristol Robotics Lab's *Human-Robot Interaction* (15 people) and *Driverless Vehicles* (5 people) research groups. I directly line-manage 4 students and early career researchers.

Supervision of graduate students and postdoctoral fellows

2018 – 2019	**2 post-docs**, **5 PhDs**, **4 MSc students**, Bristol Robotics Lab, UWE, UK
2015 – 2018	**3 PhDs**, Plymouth University, UK
2013 – 2015	**5 PhDs**, **5 MSc students**, EPFL, Switzerland
2012 – 2013	**2 MSc students**, LAAS-CNRS, France

Teaching activities

2019 –	**Associate Professor** teaching at postgraduate level, UWE, UK
2018 – 2019	**Senior Lecturer** teaching at postgraduate level, UWE, UK
2015 – 2018	**Lecturer** teaching at undergraduate & postgraduate levels (robotics fundamentals, software engineering, human-robot interaction), Plymouth University, UK
2013 – 2015	**Teaching Assistant** teaching at undergraduate level (Visual Computing), EPFL, Switzerland
2008 – 2012	**Teaching Assistant** teaching at undergraduate level (programming, databases, ontologies), INSA Toulouse, France

Contributions to the wider research community

Organisation of scientific meetings

2020 **ACM/IEEE Human–Robot Interaction conference**, 700+ participants, local chair, Cambridge, UK

2017 **ACM/IEEE Human–Robot Interaction conference**, 400+ participants, alt.HRI chair, Vienna, AT

2016 **2nd Intl. workshop on Cognitive Architecture for Social HRI**, 45 participants, programme chair, Christchurch, NZ

2014 **Intl. workshop on Simulation for HRI**, 35 participants, programme chair, Bielefeld, DE

2012 **Intl. workshop on MORSE and its applications**, 30 participants, programme chair, Toulouse, FR

2009 **Cognitive Sciences' Young Researchers Conference**, 150 participants, steering committee, Toulouse, FR

Institutional responsibilities

2019 – Full member of the EPSRC Peer Review college

2019 – Head of the Outreach cluster, Faculty of Technology and Environment, UWE, UK

2019 PhD defense committee, University of Bielefeld, DE

2019 PhD defense committee, University of Örebro, SE

2018 – HRI module co-lead, MSc level, University of the West of England, UK

2017 – 2018 Module leader, Robotics fundamentals (undergraduate level), University of Plymouth, UK

Editorial activities

2019 – Member of the Robotics, Science and System (RSS) Programme Committee

2018 – Editorial board of *Frontiers in AI and Robotics*

2017 – Member of the IJCAI Programme Committee

2015 – Member of the IEEE/ACM HRI Programme Committee

2017 – 2019 Member of the IEEE IROS Programme Committee

2017 – 2018 Member of the HAI Programme Committee

Contributions to the broader society

Policy making

2020 – **Expert Collaborator for the European Joint Research Centre** contributing to the UNICEF Guidelines for Responsible Child-Robots Interactions

2019 **Invited panel by the EU Research Executive Agency** at the 2019 MSCA AI Cluster, sharing expertise in Human-Robot Interaction

Technology transfer

2018 –	Co-I on UKRI InnovateUK projects ROBOPILOT, CAPRI, CAVForth, involving direct transfer of technology for automated verification of autonomous vehicles
2018 –	Scientific advisor for KickSum Ltd., in the frame of the EU-funded SABRE project
2018	Co-inventor on US patent US20190016213A1 on back-driveable, haptic locomotion for small robots

Selected outreach and public dissemination

2019–	Cluster Lead for STEM outreach, University of the West of England
2019–	Scientific advisor for the WeTheCurious Bristol's science Open City Lab project
2019	Hosted large media event for the Couch25K study [109]
2016–	UK & EU Robotics Weeks coordinator, University of Plymouth, University of the West of England
2015	Hosted large media event for the CoWriter study [57] (coverage by Reuters, BBC Arabic, FastCompany)
2011	'Roboscopie' Human-Robot public theater performance, Science Day'11 http://bit.ly/1LQpNWA
2008–2011	Toulouse's Cognitive Sciences Students Association, Co-chair
1997–2012	Executive Committee & Head of Educational Robotics, Planète Sciences (including coordination of the *EUROBOT* Robotic Competition)