WizUs- Part B

Lemaignan

September 21, 2023

ERC Consolidator Grant 2024
Research proposal [Part B1]

Socially-Driven Robots to Support Human-Human Interactions

WizUs

- Principal Investigator: **Dr Séverin Lemaignan**
- Host institution: **XXX**
- Duration: **60 months** (5 years)

Abstract

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 the WizUs research project, and our 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.

We 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 WizUs project.

(2) Long-term acceptance requires genuine involvement of the end-users at every step of the design process. To this end, WizUs 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.

WizUs 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 three years of field deployment of social robots in public spaces.

WizUs 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.

B1.a. Extended Synopsis of the scientific proposal

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

Over the 5 years of the 'WizUs' project, I will design and deliver a ground-breaking embodied AI for socially intelligent robots, with long-term social utility and demonstrated acceptance in the field.

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;**
- **integration of the robot's disparate perceptions into 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 these robots in two, one-year-long demonstrations in high impact, socially sensitive environments.

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. How can we ensure 'by design' that they will have a net social utility? In other words: at the age of deep neural network and large language models, what are the conditions for ensure responsible social robots?

WizUs will build the **new science required for robots to represent, reason and act in complex social environments**, with the goal of realising **a vision of social robots that enable humans and humans relationships to thrive**. WizUs is about creating the conceptual and technical frameworks required for safe and responsible social robots, intrinsically driven to foster stronger social interactions with and between humans.

WizUs main objective is, within 5 years, to design, implement and demonstrate in the real-world the AI engine of a responsible socially intelligent robot. Using the complex use-case of social isolation in elderly care centres, we will show that a social robot that implements Responsible Robotics principles, is accepted and adopted on the long term by the stakeholders, and can have a genuinely positive, long lasting, impact on the well being of older people.

This objective is underpinned by two research hypotheses: (**H1**) for end-users to ascribe social utility and engage with the robot over long periods of time (months, years), the robot has to have its own long-term internal motivation to be socially helpful – a *social teleology*.

(**H2**) Additionally, long-term acceptance also requires the genuine involvement of end-users in the shaping of the robot behaviours. As I have shown in my past research, this user engagement generate trust, feeling of ownership, and foster acceptance. Extrapolating from my laboratory result on interactive machine learning, I hypothesise that this process may lead to *mutual accomodation* and eventually long term acceptance of a robot able to continuously adapt to the need of its users.

To test and verify these two hypotheses in the real-world requires **breaking new ground in science and engineering**. Critically, we need to endow robots with a powerful way of representing and reasoning about their social environment. WizUs aims at achieving this breakthrough by developing the mathematical models underpinning the recently discovered *social embeddings*, and significantly expanding their expressiveness.

My research vision is of socially-useful robots which progressively learn to become autonomous, with the direct help and guidance of their end-users. By doing so, the stakeholders actively shape the robots' roles and behaviours, based on their actual, real-world needs and experience, while ensuring ethical behaviours.
As such, my research program is about exploring, designing and implementing novel methods for *social learning* for assistive robots. *Social learning*, in this context, means developing

machine learning techniques using *real world*, *in-context* demonstrations by the *end-users themselves* to learn task and social action policies for the robots.

The envisioned outcome is not only assistive robots with a higher degree of social autonomy, able to flexibly adapt, but also robots deeply shaped and *owned* by their users, and thus more readily accepted and adopted.

Achieving this vision requires the combination of complex robotic socio-cognitive capabilities, state-of-art machine learning techniques, and novel experimental methodologies. Specifically my research will explore and enable:

- **compact, embedding-based representations of the social and spatial context of the robot;**
- **Attention-based deep machine learning architectures to learn context-appropriate behaviour sequences for social robots;**
- **'end-user in-the-loop' data acquisition methodologies, including immersive teleoperation and progressive autonomy;**
- **the study of adoption barriers to such autonomous robots in human environments, with an initial focus on healthcare and elderly care.**

In addition, I will deliver the conceptual and ethical framework required to further support the public debate and policy making process around intelligent autonomous social robots, also through real-life demonstrations and deployments of assistive robots in high impact, complex social environments.

Closely aligned with national and European research priorities, this research program creates a excellent opportunity to reinforce INRIA and Europe as worldwide leaders in Social and Intelligent Robotics.

Framing and research objectives

AI and robots are emerging as key factors to successfully address modern societal challenges, like the ageing society or increasing social isolation. In this context, how to ensure *by design* that social robots have a positive social impact? This question is the backbone of my research project, and my research vision is to **create within 5 to 10 years socially-intelligent and responsible robots, that (1) will have recognised social utility, and (2) will see long-term acceptance by their users**.

This translates into three overarching, long-term research questions:

- What are the public expectations with respect to the role of social robots, and how can we **collaboratively design autonomous**, yet **responsible, beneficial, socially acceptable robots**?
- What are the conceptual, algorithmic and technical prerequisites to design and implement such an autonomous & responsible robots? in particular, what social context understanding and (machine) learning architectures are required to **enable long-term autonomy** and, eventually, **engagement** between a robot and its end-users?
- What are the conditions and methodologies enabling large scale data acquisition of **real world, user-driven robots behaviours**? How to then train robots to become **progressively autonomous**? And ultimately, how to balance **autonomy** of the robot with the necessary **behaviour transparency** and **human oversight**?

From these questions, I derive the following four objectives that are the guiding principles of my research program, both in the short term, and at a 10-15 years horizon:

Methodology

Actual utility and long-term acceptance requires genuine involvement of the end-users at every step of the design process. This is at odds with the common, engineering-centered practise of first developing robots and algorithms *ex-vivo*, in lab, and then placing a (semi) final product in the hands of the users, hoping for adoption. Adoption, however, is the result of a long process of *mutual modelling* [50], where the social role of the robot is slowly constructed from its real world, in-context usage.

This foundational insight requires the conceptual framing and development of new research methodologies. I have started to explore these questions in some of my previous work [56, 67, 69], and my research program aims at significantly developing this line of research to tackle more complex, long-term application

domains.

One of the major challenge arising with more complex application domains is however the combinatorial growth of the problem space. Indeed, none of the current control paradigms or cognitive modelling techniques are able to successfully predict context- and task-appropriate sequences of behaviours for socially-useful autonomous robots.

My research programme aims at tackling this challenge with two key insights: (1) the representation complexity of the social and spatial environment of the robot can be dramatically reduced by treating it as an *embedding* problem and applying modern machine learning techniques (like GANs) to design and compute *social embeddings*; (2) modern transformers and attention-based [59] machine learning architectures have demonstrated long-term modelling capabilities on complex language domains [47] that could in principle be equally applied to action sequence generation for robots. Initial explorations of this second insight have started to emerge [4, 60], but none of these early research efforts consider the complexity of human interactions.

My research program will research how these insights can be effectively operationalized, with a scientifically ambitious and highly technical work program. It includes basic research and conceptual framing; extensive, beyond-state-of-art, technical developments; and an ambitious experimental program, centered on long-term 'user in-the-loop' data collection via field deployments of social robots in public spaces – and primarily in the healthcare and elderly care environment.

Work plan outlook

My research program could begin rapidly, using publicly available resources, including machine learning architectures like Transformers, combined with open-source pre-trained Large Language Model backbones; and state-of-art HRI tools like ROS4HRI [44] to represent in real-time the social environment of the robot. While long and complex data collection campaigns would have to be organised, and training infrastructure would need to be designed, I expect initial results in the first 3 to 5 years.

This is also a long-term vision: on the one hand, the rapid pace of progress of technology (novel deep machine learning architectures, novel HRI tools for human and scene understanding) continuously opens novel investigation venues; one the other hand, the success of my research vision hinges on real-world, long-term experimental work: deploying robots in the healthcare sector, creating the conditions for adoption by the end-users, running long-term deployments with the end-users are long terms aims ,... these research activities will take place over long period of time.

Importance and impact

My research program has the potential to be groundbreaking: until now, autonomous social robots have had little real world success. Experiments and deployments have been mostly limited to constrained application domains, where rigid action policies (scripts, task planners) could be sufficient. State-of-art robots however fail to handle the complexity and unpredictability of real world environments (like the ones encountered in the healthcare domain). In addition, these systems see poor field adoption due to several factors including difficulty of use, wrong expectations, perceived complexity.

The novel paradigm that I will develop and deploy as a Directeur de Recherche at INRIA addresses both this limitations. By using state-of-the-art machine learning techniques – with powerful abilities to adapt to unknown context – combined with a novel 'user-in-the-loop' approach to data collection and behaviour shaping, I believe we can overcome both challenges: real-world autonomy, and adoption by the end-users.

This research program is also hugely important: as socially assistive robots quickly develop, it is critical to equip ourselves with a deeper understanding and intellectual framing of what social robots *can* and *should* be, paving the way for their much broader adoption in the coming years: as a Directeur de Recherche, I will actively contribute to this aim, by leading the design and implementation of socially-intelligent robots that are socially useful, acceptable in the long-term, and ethically responsible, but also by furthering my engagement to interdisciplinary work, and broad engagement with the society and policy makers.

I frame these hypotheses with the idea of **robot-supported human-human interactions**, a novel conceptual framework to 'think' the future human-robot interactions. I will co-construct this framework through large scale public engagement: for a whole year, I will deploy the WizUs robot within the City Lab of Bristol's science centre *We The Curious*, relinquishing the control of the robot to the visitors themselves. Tasked with remotely operating the robot to assist fellow visitors, I will accompany them in 'inventing by doing' a new grammar of social interactions: 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 had experienced – and co-designed – how robots should interact with humans in a positive, helpful way, and 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, two further large scale experiments will take place over the course of the project: a one-year experiment in one of Bristol's Special Education Needs (SEN) school, helping 250+ children with psycho-social impairments to develop their social skills; a second one-year experiment at the Bristol's children hospital, where the robot will join one of the wards where 30+ children with long-term conditions stay for months, and engage with the children into playful social activities: telling stories, triggering group activities with other children, providing additional social presence. In both these experiments, the robot behaviours will be co-designed with, and learnt from the end-users themselves: nurses, teachers, parents, and where possible, the children themselves.

Importantly, WizUs focuses specifically on the AI engine of the robot: I will use an existing robotic platform (PAL Robotics TIAGo Pro, 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 tools to override the robot 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. To this end, the robot will have ground-breaking perception and modelling capabilities to represent the current social situation (the focus of WP2), coupled with an innovative cognitive architecture designed to combine internal social drives with domain-specific action policies learnt from the end-users (WP4). The robot actions themselves are designed to be limited to non-verbal communication mechanisms: non-verbal utterances using sounds, gaze, joint attention, expressive motions. In WP3, my team will work with a choreographer and a sound expert to create a new grammar of expressive motions, combined with a novel modality based on *soundscapes*: sound landscapes that the robot can modulate to influence the mood of the social environment (calm, excited, worried, etc.).

As a ground-breaking project, WizUs will assert and reinforce the European leadership in AI and intelligent robotics, in line with the EU's strong societal values: by developing socially responsible AI that guarantees, by design, long-term benefits to the society. The very first task T1.1, spread over the first 4 years of the project, specifically addresses and frames the ethical underpinnings of social robots, and delivers the guidelines that we need to inform our future policies on social robotics. Combined with beyond-state-of-the-art technological developments, **the WizUs research programme will provide a major contribution in securing a safe and responsible digital future in Europe**.

Finally, WizUs is also about asserting and reinforcing the European leadership in AI and intelligent robotics, in line with EU strong societal values: a socially responsible AI, that guarantees, by design, long-term benefits to the society. This requires leading major technological advances; leading the development of the conceptual framework around socially intelligent robots that we need to inform future policy making; but also **a strong leadership to meaningfully involve the public at large in the design of these technologies**. Through its objectives and methodology, **WizUs will have a major contribution to building this capacity in Europe**.

Overview of the WizUs work programme

Socially intelligent robots require unique, beyond state-of-the-art, capabilities to *(1)* understand the social interactions (social situation awareness), *(2)* autonomously decide the best course of action for short-term and longer-term social influence, and *(3)* perform the appropriate social actions and exert said influence in an appropriate, responsible manner. Not only the required technology is itself beyond state-of-the-art (and will be researched and integrated in WP2, WP3 and WP4), but the interplay between technology, socio-cognitive psychology, privacy and ethics is only starting to be researched and understood. WizUs offers an strong vision and an ambitious, evidenced-based, methodology to significantly advance our understanding of this multi-faceted problem.

Over the course of 5 years, I will investigate hypotheses H1 and H2 by addressing the following research objectives:

- **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?

Building on the extensive body of work on Responsible AI, I will investigate the basic principles of responsible robot-mediated social interactions, that must form the foundations of a socially useful robot, accepted and used in the long run. Using user-centred design and participatory design methodologies, I will identify the determinants and parameters of a responsible social intervention, performed by a socially-driven robot, and formalise them in practical principles.

- **O2: physical-social representation and reasoning** To effectively and responsibly interact with its environment, the robot must first build a comprehensive and continously updated model, from its spatial and physical configuration, to its social dynamics. I will design and develop a novel cognitive capability of artificial *social situation assessment* to enable the robot to represent real-time social dynamics in its environment. I will achieve this breakthrough by combining existing model-based approaches **TODO: refs** (including my recent research on social state modeling **TODO: refs**, with the expressive power of the new *social embeddings* that I have recently introduced.
- **O3: goal-driven, responsible decision making** I aim to create robot behaviours that are perceived as purposeful and intentional (long-term goals), while being shaped by a user-created and user-controlled action policy. I will integrate long-term social goals, arising from the interaction principles of **O1**, with the social modeling capability of **O2**, into a principled, goal-driven cognitive architecture, with responsible AI guarantees. 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 attention-based machine learning.

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

  Second, that human-in-the-loop machine learning can be used to ensure an additional layer of human oversight and a level of behavioural transparency. Human-in-the-loop reinforcement learning – as implemented in the SPARC approach that I have developed with my students and already used in complex social environments [52, 54, 68] – relies on an end-user 'teacher'. This teacher initially fully controls the robot (via teleoperation) while it learns the action policy, and then progressively relinquishes control up to a point where the robot is effectively autonomous. As I previsouly argued in Senft et al. [54], this approach leads to increased control and ownership of the system, and as a result, increased trust from the end-users.
- **O4: 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 *ecosystems* that are sufficiently complex and open to explore novel social interactions. My objective is also to show that this real-world deployment can be successfully driven by the 'end-to-end' involvement of all the end-users and stakeholders: from defining the robot's role, from the different perspective of each end-user, to actually designing and 'teaching' the robot what to do.

These objectives are investigated across five work-packages.

WP1: Framing robot-supported human-human interaction

WP1 aims at establishing the conceptual and ethical framework around the idea of *robot-supported human-human interactions*. It does so by co-creating patterns of interaction and norms with the general public, using a unique combination of ethnographic observations and 'crowd-sourced' interaction patterns.

Main outcomes: A theoretical framework to 'think' the role of social robots and guidelines to inform policy making (including ethical implications); a set of operational & co-created interaction principles; a large dataset of social human-robot interactions

Timeframe: Y1-Y3; one senior post-doc (PD1) with background in sociology of technology.

T1.1 – Conceptual framing and ethics of robot-supported social interactions

The first task in WP1 is to research and define the conceptual framework 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. In addition, four two-days workshops with the WizUs Ethics Advisory Board, spread over the duration of the project, will act as ethical milestones.

T1.2 – Crowd-sourced patterns of robot-supported social interactions The conceptual framework identified in T1.1 is translated into a set of *interaction design principles* and *determinants* that will together form a set of requirements and objectives for the socio-cognitive capabilities and architecture developed in WP2 and WP4.

In order to anchor T1.2 into the reality and complexity of human social interactions, and to involve the society at large into the design of these patterns and norms, I will embed one WizUs robot in the Bristol Science Centre WeTheCurious for a whole year (Y2). With the help of a researcher, the visitors will be guided into tele-operating the robot to assist other visitors, and, by doing so, co-design what a good robot helper should be. This will generate the quantitative and qualitative data to inform questions like 'what role for the robot?', 'when to intervene?', 'what are the effective and acceptable social influence techniques?'. It will also be a unique example of crowd-sourcing at a large scale, with the general public, the interaction design of social robots. The generated dataset will also be used as data source in WP2 and WP3.

Specific resources The Bristol's Science Centre is fully committed to the project, will include WizUs in its official programme of activities, and will provide in-kind training for the WizUs researcher based at the centre.

WP2: Real-world Social Situation Assessment

In WP2, the project addresses the key scientific and technical pre-requisites to effectively deliver WP4's cognitive architecture; namely the perception and modeling of the spatio-temporal and social environment of the robot. This includes spatial characteristics (proxemics; group dynamics; complex, dynamic attentional mechanisms); psycho-social determinants (social roles and hierarchies; social groups; mental modelling; anthropomorphic ascriptions); temporal characteristics (effects of novelty; dynamics of anthropomorphism and mental ascriptions; group dynamics). I have investigated many of these socio-cognitive capabilities in isolation (Table 1.1), and this WP is about *integrating* them into a coherent perceptual subsystem, significantly extending the state-of-the-art [38, 3].

Main outcomes: A complete pipeline for spatio-temporal and social situation assessment, build as open-source ROS nodes, and able to map in real time the physical and social environment of the robot.

Timeframe: Y1-Y4; one post-doc (PD2) in social signal processing/machine learning/cognitive modelling.

T2.1 – Hybrid situation assessment and knowledge representation This task builds the foundational spatio-temporal and symbolic perception and representation system for the robot. It will integrate the state-of-the-art in spatio-temporal situation assessment that I have previously developed [37, 51] with recent advances in data-driven semantic labelling (for instance, using 4D convolution nets like MinkowskiNet [5]), and a symbolic knowledge base (like my own ontology-based one [32]) in order to create a coherent system of representations for the cognitive architecture of the robot.

T2.2 – Multi-modal human model This task focuses on the processing and modelling of social signals, extending existing techniques, both model-based (eg [16, 30]) and data-driven based (eg [1]). This task goes beyond the state-of-the-art by looking specifically at resolving highly dynamical signals (like gaze saccades and micro facial expressions). Required datasets will be drawn from my previous work [34], as well as from the project experiments (T1.2, T5.1, T5.2).

T2.3 – Interaction and group dynamics Building on T2.2, T2.3 investigates the automatic understanding and modelling of group-level social interactions [58], including f-formations [42], sociograms (as done in [13] for instance), and inter-personal affordances [48]. This task builds on literature on social dynamics analysis (eg [8, 21, 43]) to apply it to real-time social assessment by a robot, itself embedded into the interaction.

T2.4 – Integrated model of the social environment The integration of the social cues from T2.2 and T2.3 results in a socio-cognitive model of the social environment of the robot, that effectively extends the representation capabilities of T2.1 to the social sphere. The result of T2.4 is an AI module that implements a full social assessment pipeline, from social signal perception to higher-level socio-cognitive constructs. T2.4 also includes a focused experimental programme (based on the protocols designed by Frith and Happé [12], that I introduce in [27]) to demonstrate in isolation the resulting socio-cognitive capabilities.

WP3: Generative social behaviours

Mirroring WP2's focus on understanding the social interactions, WP3 addresses the question of social behaviour *production*: how to create natural, non-repetitive behaviours, engaging over a sustained period of

time. The robot behaviours will be exclusively non-verbal (non-verbal utterances, gaze, joint attention, facial expressions and expressive motions), and will include soundscapes as a novel non-verbal interaction modality.

Main outcomes: A new method to automatically design complex and non-repetitive social behaviours, with a focus on non-verbal communication; research on soundscapes as a novel non-verbal modality for human-robot interaction.

Timeframe: Y2-Y5; one post-doc (PD3) in machine learning/learning from demonstration.

T3.1 – Behavioural baseline T3.1 establishes a baseline for behaviour generation, by surveying and implementing the current state of the art (behaviours library, activity switching [6]). This baseline will enable early in-situ experimental deployments, while also providing a comparison point for T3.2.

T3.2 – Generative neural network for social behaviour production WizUs aims 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 [41, 57] (with training data created with an expert choreographer); (2) interactive machine learning in high-dimensional input/output spaces, where I have shown with my students promising results for generating complex social behaviours [54, 66] that fully involve the end-users [70]. Modulating (1) with the learnt features of (2), I target a breakthrough in robots' social behaviours generation: the generation of non-repetitive, socially congruent and transparent social behaviours (including gestures but also gazing behaviours and facial expressions).

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 interventions [15]. The soundscapes that I will create are 'owned' by the robot, that can manipulate ithem 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 user.

Specific resource: these soundscapes will be co-designed with Dr. Dave Meckin, an expert on sound design for vulnerable children, who also works at the host institution.

WP4: Goal-driven socio-cognitive architecture

In WP4, I will design a novel socio-cognitive architecture for the social robots, and implement it on the IIT R1 robot. WP4 will integrate together the modeling capabilities and behaviour production developed in WP2 and WP3, with a dual action policy: a policy driven by a social teleology (eg an artificial intrinsic motivation to act socially), and a policy learned through human-in-the-loop machine learning. This WP is high-risk/high-gain: while sustaining long-term engagement in a principled way remains one of major scientific challenge in social robotics [17], the WP suggests a very novel approach to goal-driven socio-cognitive architectures, with the potential of unlocking long-term social engagement by endowing the robot with its own intentionality [65], while maintaining human oversight.

Main outcomes: An integrated cognitive architecture for social robots, driven by both long-term social goals, and machine-learnt action policies; a reference open-source implementation, enabling long-term autonomy on the IIT R1 robot.

Timeframe: Y1-Y5; one post-doc (PD4) in cognitive robotics; one PhD student (PHD1).

T4.1 – A social teleology for robots The idea of a *teleological* (ie goal-driven) robot architecture for social interaction is very novel (existing literature on teleological robots only focuses on simple cognitive systems [46, 11]). This task designs and implements such an architecture on the R1 robot. It first identifies from the interaction patterns and determinants uncovered in T1.2 *interaction principles*, that are then mapped into *long-term interaction goals*, capable of driving the robot actions over a period of time.

T4.2 – Learning from humans to achieve 'by-design' responsible & trustworthy AI Building on my recent results on human-in-the-loop social learning [52, 54, 66], this task implements the mechanics to allow human end-users to progressively teach the robot a domain-specific social policy. This task also qualitatively researches 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, resulting in explainable behaviours for the end-users.

T4.3 – Integrating a socially-driven architecture for long-term interaction Building on my previous work on cognitive architecture [38], this task brings together, in a principled manner, the perceptual (WP2) and behavioural (WP3) capabilities of the robot, as well as the social policies created in T4.1 and T4.2.

T4.3 will specifically look at long term autonomy, including long-term social goals, cognitive redundancy, and behavioural complexity.

T4.3 will also develop the arbitration mechanism that combines the robot's social teleology (T4.1) with the human-taught action policy (T4.2). This arbitration mechanism will build on research on reinforcement learning for experience transfer [39] that enables the re-assessment of a policy (here, our intrinsic motivation) based on previous experience (here, the human-taught policy).

WP5: Experimental programme: long-term deployments in sensitive social spaces

Finally, WP5 aims at convincingly demonstrating the importance and positive impact that socially-driven, socially-responsible robotics may have. The experimental work of WizUs will be organised around two ambitious long-term studies, in complex, real-world environments: a network of special educative needs (SEN) schools, and the Bristol Children's Hospital.

These environments also put the project in the unique position of actually delivering high societal impact: I anticipate 30+ hospitalised children, and 250 SEN-educated children to directly benefit from the project, exploring how robots can have a net social utility, while being accepted as effetive tools by field practitioners. Both these deployments will take place within the strict ethical framework established in T1.1.

Main outcomes: Two long-term deployments of a social robot in real-world, high impact environments demonstrating long-term acceptance and social utility; large (anonymous) datasets of complex, real-world human-robot interactions.

Timeframe: Y3-Y5; one post-doc (PD4, shared with WP4).

T5.1 – A robot companion to support physical, mental and social well-being in SEN schools This task aims at demonstrating robot-supported social interventions within Bristol's network of SEN schools. During a one-year period (Y3), the robot will be based in schools, with interventions co-designed with the teachers, the parents and the students, both through preliminary focus groups and in-situ machine learning.

The envisioned interventions include: initiating group games; enquiring students about their well-being; co-teaching material with teachers; fostering interaction situations between the children.

Specific resources: The task will be supported by local SEN researcher Dr. Nigel Newbutt, who has a long track record and on-going research partnerships with Bristol's special needs schools.

T5.2 – A robot companion to support isolated children during their hospital stay Over the course of this second, one-year long (Y4) experiment, my team will deploy one WizUs robot in the long-term condition paediatric ward at the Bristol Children's Hospital. Using a *mutual shaping* approach [70] to design the role of the robot with the different stakeholders (nurses, doctors, parents, children), I will 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 amongst children, within the ward.

Specific resources: Several preparatory meetings already took place with the head of the hospital education service J. Bowyer, who will support the project, granting access to the long-term conditions ward, and sponsoring the project through the hospital-specific ethics process.

Capacity of the Principal Investigator to deliver on the work programme

While the project is ambitious, I am in a unique position to deliver on the WizUs work plan. I already have established international recognition in human-robot interaction and have likewise demonstrated strong leadership by leading research teams in three different institutions (see Sections B1.b and B1.c below). Importantly, as illustrated in Table 1.1, the breadth of my interdisciplinary research covers the scientific expertise required by the project, providing me with a unique overall perspective and understanding of the domain. I am also a technology expert, with major software and hardware contributions to the robotic community (see Section B1.c). As such, I have a excellent grasp of the technical feasibility of the proposed work.

The project is ambitious, with an experimental programme that goes significantly beyond the state of the art. It will provide a lasting scientific and technical legacy, that extends well beyond the end of the fellowship. As a high-risk/high-gain project, WizUs will also be a powerful enabler: by the end of the fellowship, I will have established myself as a world-leader in the emerging field of socially-driven, responsible autonomous robots, significantly reinforcing the European capacity in this critical field for our digital future.

Table 1.1: PI's domains of expertise relevant to the WizUs project

Psycho-social underpinnings of HRI	
human factors	anthropomorphism [28], cognitive correlates [29], social influence [67]
trust, engagement, social presence	[10, 35, 9, 19]
theory of mind	perspective taking [49, 64], social mutual modelling [27, 7]
Social signal processing	
non-verbal behaviours	attention [30], child-child dataset [34], internal state decoding [1]
verbal interactions	speech recognition [23], dialogue grounding [36]
Behaviour generation	
social behaviours	[24], verbal interactions [62, 63], physical interactions [14]
interactive reinforcement learning	[55, 52, 54]
Socio-cognitive architectures	
architecture design	[38, 3, 26, 25, 40]
knowledge representation	ontologies [32, 33]
spatio-temporal modelling	object detection [61], physics-aware situation assessment [37, 51]
Fieldwork in HRI	in classrooms [18, 31, 20, 2, 22, 53], at home [45]

Bibliography

[1] M. Bartlett, C. 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 Robotics and AI* (2019).

[2] 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.

[3] 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.

[4] A. Brohan et al. *RT-1: Robotics Transformer for Real-World Control at Scale*. 2022. DOI: 10.48550/ARXIV.2212.06817. URL: https://arxiv.org/abs/2212.06817.

[5] C. Choy, J. Gwak, and S. Savarese. "4D spatio-temporal convNets: Minkowski convolutional neural networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 3075–3084.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] H. Gunes and B. Schüller. "Automatic Analysis of Social Emotions". In: Cambridge University Press, 2017, p. 213.

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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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).

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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).

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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. DOI: 10.1007/s12369-011-0123-x. URL: http://dx.doi.org/10.1007/s12369-011-0123-x.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] Y. Mohamed and S. Lemaignan. "ROS for Human-Robot Interaction". In: *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2021. DOI: 10.1109/IROS51168.2021.9636816.

[45] 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.

[46] 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.

[47] L. Ouyang et al. *Training language models to follow instructions with human feedback*. 2022. DOI: 10.48550/ARXIV.2203.02155. URL: https://arxiv.org/abs/2203.02155.

[48] 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.

[49] 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.

[50] S. Šabanović. "Robots in Society, Society in Robots". en. In: *International Journal of Social Robotics* 2.4 (Dec. 2010), pp. 439–450. ISSN: 1875-4791, 1875-4805. DOI: 10.1007/s12369-010-0066-7.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[60] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor. *ChatGPT for Robotics: Design Principles and Model Abilities*. Tech. rep. Microsoft, 2023.

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

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

[63] 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.

[64] 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.

[65] E. Wiese, G. Metta, and A. Wykowska. "Robots as intentional agents: using neuroscientific methods to make robots appear more social". In: *Frontiers in psychology* 8 (2017), p. 1663.

[66] 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.

[67] 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.

[68] 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.

[69] K. Winkle, E. Senft, and S. Lemaignan. "LEADOR: A Method for End-To-End Participatory Design of Autonomous Social Robots". In: *FrontiersIn AI and Robotics* (2021). DOI: 10.3389/frobt.2021.704119.

[70] 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.

B1.b Curriculum-vitae and track record
Dr. Séverin Lemaignan

Personal details

ORCID: 0000-0002-3391-8876
Nationality: French
Date of birth: 17 Jan 1983 (37 years old)
academia.skadge.org – twitter.com/skadge

Education and key qualifications

2008 – 2012 **Joint German-French PhD in Cognitive Robotics**
LAAS-CNRS, France / Technical University of Munich, Germany
Supervisors: Pr. Rachid Alami, CNRS; Pr. Michael Beetz, TUM

2004 – 2005 **MSc Artificial Intelligence for Learning Technologies**
University Paris V, France

2002 – 2002 **Joint German-French MSc of Engineering**
Karlsruhe Institute of Technology, Germany / ENSAM ParisTech, France

Current position

2021 – **Head of Social Robotics and HRI Research**
PAL Robotics, Barcelona, Spain
Head of the Human-Robot Interaction research and enginerring group.

Previous positions

2019 – 2021 **Associate Professor in Social Robotics and Artificial Intelligence**
Bristol Robotics Laboratory, University of the West of England, United Kingdom
Head of the Human-Robot Interaction research group; Head of the Driverless Vehicle research group. Directly managing 20+ students and early career researchers.

2018 – 2019 **Senior Research Fellow in Robotics and AI**
Bristol Robotics Laboratory, University of the West of England, United Kingdom

2017 – 2018 **Lecturer in Robotics**
Plymouth University, Plymouth, United Kingdom

2015 – 2017 **EU Marie Skłodowska-Curie Post-doctoral fellow**
Plymouth University, Plymouth, United Kingdom
Development and Implementation of a Theory of Mind for robots

2013 – 2015 **Post-doctoral fellow**
CHILI, EPFL, Lausanne, Switzerland
Interaction with Robots in Learning Environments – Supervision of the robotic group

2012 – 2013 **Post-doctoral fellow**
LAAS-CNRS, Toulouse, France
Spatial and Temporal Reasoning for Cognitive Robotic Architectures

2006 – 2007 **Research Engineer**
INRIA, Paris, France
Development of semantic-aware control architectures for autonomous vehicles

RESEARCH ACHIEVEMENTS AND PEER RECOGNITION

Research achievements

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

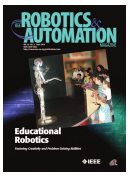
A novel representation technique to efficiently represent multiple parallel states of the world, including imaginary ones. This ability is critical to represent spatio-temporal predictions, and to create models of other agents' representations. **[principal investigator]**

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 rapidely thanks to an innovative input feature selection mechanism.
[student supervisor; 22 citations]

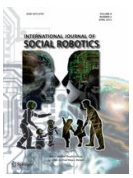
Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.
Artificial Cognition for Social Human-Robot Interaction: An Implementation
Artificial Intelligence 2017

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; 140 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 Mod-

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

Peer recognition

2023	**chosen as General Chair of the HRI'25 conference**
2020	**invited to the international HRI Steering Committee**
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
2012	**Best PhD in Robotics 2012** award, CNRS, France

Additional Information

Career breaks, diverse career paths and major life events

Other contributions to the research community

B2.a State-of-the-art and objectives

TODO: TARGET PAGE COUNT: 3 pages: state of art + vision; 2 pages: methodology overview; 5 pages: WPs; 3 pages: ethics + risks; 2 pages: resources

TODO: as a reference: DECRESIM project: 4 pages on B2.a State of art and objectives; B2.b 7 pages on WPs + 2 pages on risk assessment

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 [81]. 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 [78].

Indeed, understanding *why* these robots have failed, is one of the active debate within the Human-Robot Interaction community [34], with only a handful of qualitative studies on this question [21, 29]. 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 [44]; the poor management of expectations, where user over-ascribe cognitive capabilities that do not match the reality. The community agrees however that the crux of the issue is achieving long-term social engagement [85, 34]

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 [38, 42], with a number of studies involving social robots deployed in real-world settings (for instance in schools [41, 80, 45, 18], homes [29] and care centres [32, 83]) 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 [32] or in the H2020 MuMMer [33] and FP7 Spencer [77] projects), or a narrow, well-specified complex task (for instance, supporting exercising in a gym, as I did in [83]). However, general purpose, long-term interaction is still an open question.

Social robotics and vulnerable children

Application of social robotics to vulnerable children (either hospitalised or suffering cognitive and/or motor impairments) is an active field of research. This reflects both the measured efficacy of robot-based intervention, and the perceived need for additional support for these populations. Several European projects have looked at these populations, for instance the FP7 Aliz-e and H2020 DREAM projects: in Aliz-e, [8, 18] report on how a social robot can support long-term engagement with diabetic children in hospital (noting however the rather "crude" nature of the created social interactions, and the limited autonomy of the robot).

A significant body of literature also exist on robotics and autism (see for instance review by Pennisi et al. [58]). This specific domain has been a fruitful terrain to explore specific aspects of social cognition in robotics (for instance, related to the Theory of Mind or to the processing of emotions. See my review in [43]). This research draws on the extensive prior research in experimental cognitive sciences on autism (for instance [6, 27]), and the focused experimental programme of WP2 will specifically draw from this body of prior work to evidence the newly developed social modeling capabilities of the robot.

The experimental programme of WizUs will take place over two years, first in Special Education Needs schools (WP5.1), then in a paediatric ward at the Bristol's Children's Hospital (WP5.2). For both these

studies, I will build on the experience learnt from previous studies in similar environments, with the novel contributions being both the very long-term interventions (one year each), and the user-centered methodology that I describe in the following sections.

State of art in mobile robotic platforms for social interactions

Looking specifically at human-sized mobile manipulators with advanced social features, the choice of robotic platforms is in effect limited. Table 3.1 compares the two leading mobile social robots available on the market today (PAL TiaGo and SoftBank Pepper), along with the new R1 platform developed by the Italian Institute of Technology (note that the Fetch Mobile Manipulator has been omitted, as it is functionally similar to TiaGo).

While not yet on the market, the IIT has offered early access to the platform for this project.

Table 3.1: Comparison of IIT R1 robot with PAL TiaGo and Softbank Pepper. R1 has been chosen for WizUs for being the only mobile dual manipulator with good navigation and advanced social interaction capabilities.

	PAL TiaGo	Softbank Pepper	**IIT R1**
Social features	Poor (non-expressive head)	Expressive, yet fixed, face; limited gaze; approachable	Expressive face [40]; artificial skin for touch-based interactions; approachable
Perception	Medium (RGB-D camera; laser scanner; no microphone)	Medium (RGB-D; simple mic array; poor laser scanner)	Good (RGB-D; simple mic array; laser scanner)
Navigation	Good (however, limited agility due to large footprint)	Poor (weak localisation capabilities)	Good (high agility due to Segway-like self-balancing)
Safety	Medium (heavy robot; large footprint; non-compliant arm)	Medium (smaller footprint; safe arms; limited stability)	Good (smaller footprint; safe arms; dynamic stability)
Manipulation capabilities	Medium (non-anthropomorphic gripper; single arm)	Limited (poor gripper with low payload; dual arm)	Good (anthropomorphic gripper; pressure sensors; dual arm; 1.5kg payload)
Suitability for care environments	Poor (relatively large, difficult to clear)	Good (smaller footprint, easy to clean)	Good (smaller footprint, easy to clean)

WizUs aim and objectives: Responsible robots for long-term social engagement

The overall aim of the WizUs project is to **create, sustain and better understand the dynamics of responsible long-term social human-robot interactions**.

This broad aim translates in three key research questions that we seek to address over the course of the project:

RQ1: what are the public expectations with respect to the role of social robots? Can we collectively design principles ensuring safe, beneficial, socially acceptable robots?

RQ2: what AI is required to sustain long-term engagement between end-users and a robot? In particular, how to provide a robot with an understanding of its own social environment? How to create behaviours that are not repetitive or overly predictable?

RQ3: what new ethical questions are raised by long-term social interaction with an artificial agent? In particular, how to balance autonomy of the robot with behaviour transparency and human oversight?

The WizUs objectives are built around these three questions.

O1: conceptual framing 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. I will answer questions like: What should motivate the robot to step in and attempt to help? What social norms are applicable to the robot behaviours? 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. This objective aims at addressing RQ1, and is realised in WP1.

O2: real-time social modeling I will significantly extend and integrate the current state-of-art in spatio-temporal modeling (so-called *situation assessment*) with recent research in social state modeling, to create

the novel cognitive capability of artificial *social situation assessment*, enabling the robot to represent in real-time the social dynamics of its environment. This objective addresses one part of RQ2, and is investigated in WP2.

O3: congruent social behaviours production Using the state-of-the-art in generative neural networks, combined with data acquired from an expert choreographer, I will create a novel way of producing non-repetitive, socially-congruent, expressive motions. 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. This objective addresses another part of RQ2, and is the focus of WP3.

O4: embodied AI breakthrough 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. This will result in 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 specifically test two hypotheses: first, I hypothesise 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, I hypothesise 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 [63, 64, 83] – relies on a end-user 'teacher', initially fully controlling the robot (teleoperation) while it learns the action policy, and then progressively relinquishing control up to a point where the robot is effectively autonomous. As argued in [64], this approach leads to increased control and ownership of the system, and as a result, increased trust.

This addresses RQ2 and RQ3; however it also raise an 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'**: The design of a policy arbitration mechanism that preserve the robot's long-term intentional behaviour, while effectively guaranteeing human control, ownership and oversight. **O4** and **O4'** are addressed in WP4.

O5: ambitious field research Finally, the last objective of the WizUs project is to demonstrate the effectiveness of my approach in complex, real world conditions. This means deploying the WizUs robots into existing social eco-systems that are sufficiently complex, yet open to explore novel social interactions. My objective is also to show that this real world deployment can be successfully driven by 'end-to-end' involvement of all the end-users and stakeholders: from defining the robot role, from the different perspective of each of the end-users, to actually designing and 'teaching' the robot what to do. This is the focus of WP5.

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

Interdisciplinary nature of the research programme

WizUs 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, **WizUs builds bridges across multiple disciplinary boundaries**.

WizUs 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 WizUs 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 WizUs

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.

B2.b Methodology

Workpackages overview and interrelations

The four research questions previously listed are addressed across five work-packages: **WP1** is dedicated to the conceptual framing of the project (R1) and the identification of interaction principles; **WP2** extracts from these principles the set of requirements in term of socio-cognitive capabilities for the robot (R3), and implement them; in parallel to WP2, **WP3** looks at how social robots can generate congruent social behaviours (R3); **WP4** transposes the conceptual framework of WP1 into a principled cognitive architecture and integrates together the cognitive functions of WP2 and WP3 (R2);and **WP5** organises the experimental fieldwork that demonstrates the WizUs approach in ambitious and complementary real-world situations (R4).

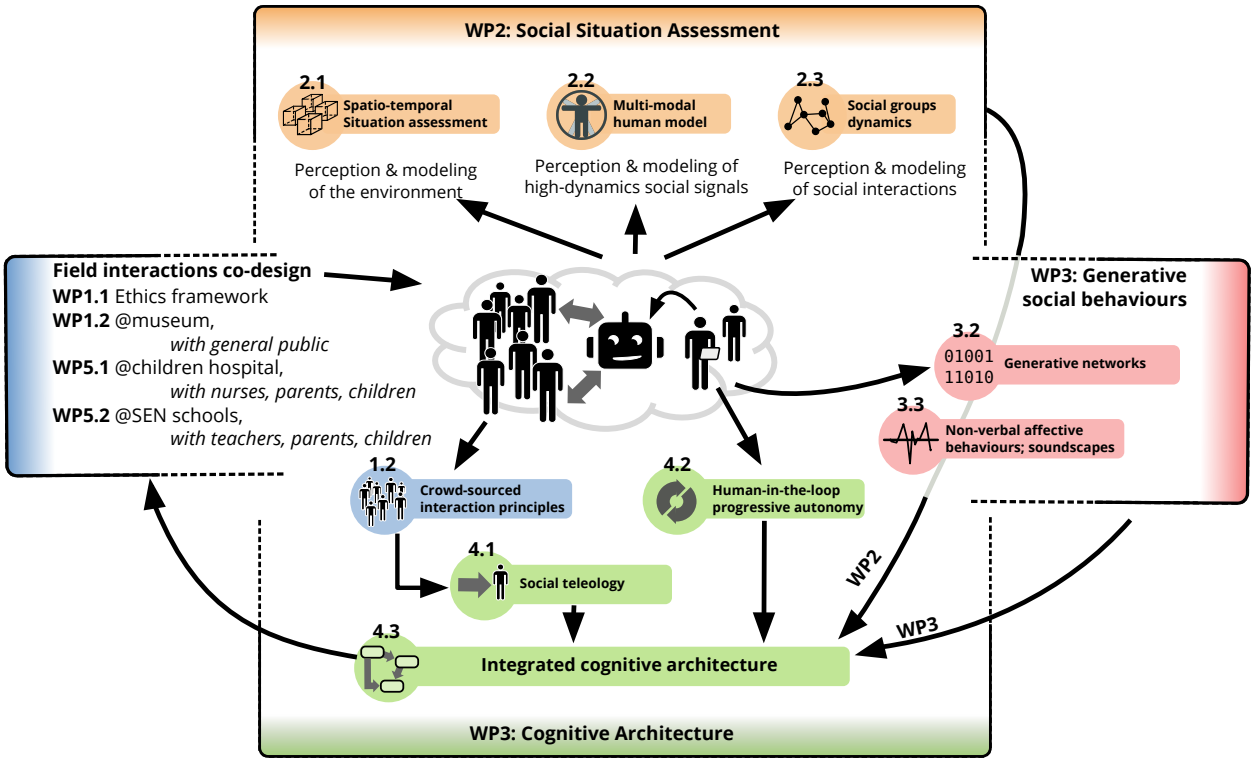


Figure 4.1: Overview of the workpackages and tasks, and tasks inter-relations.

More specifically, Figure 4.1 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.

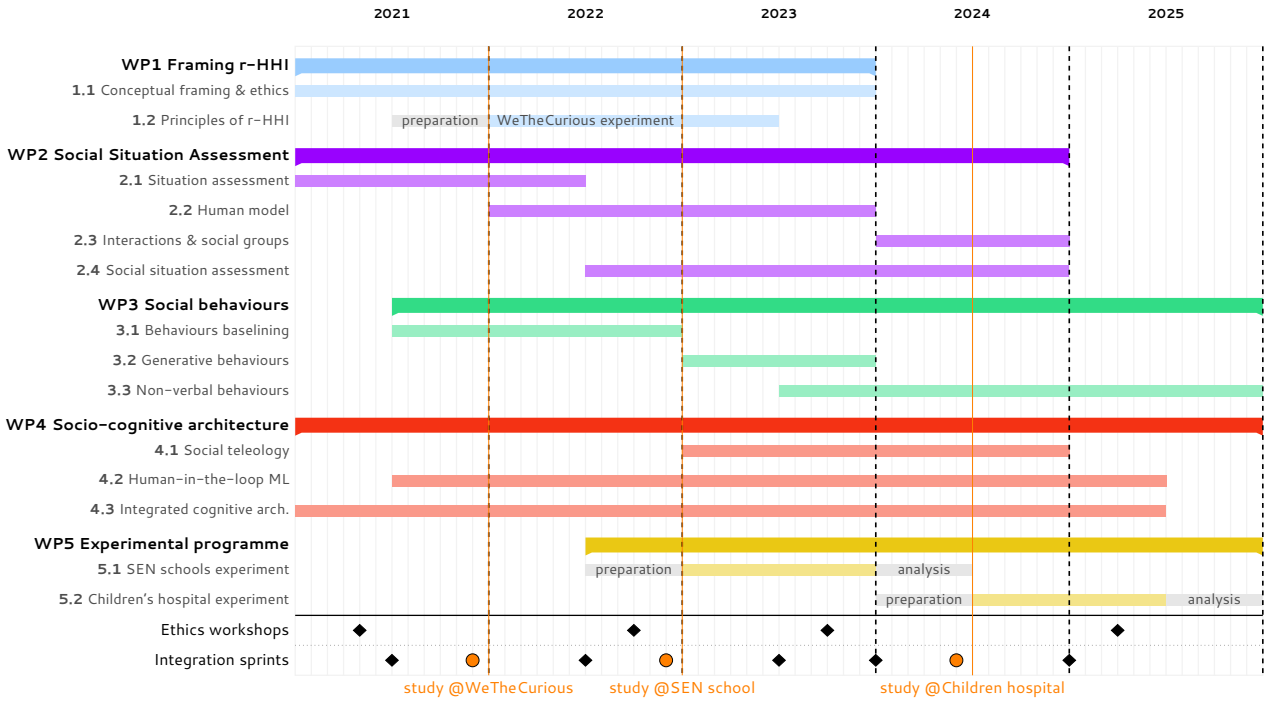
In addition to the integration of WP2 and WP3 capabilities, WP4 is also researching and developing the socio-cognitive drives of the architecture. They come both from T1.2 (as previously mentioned), and human-in-the-loop/public-in-the-loop machine learning (T4.2). T4.2, in particular, is tighly connected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

Integration sprints

WizUs is a complex project, with numerous interdependencies between tasks. To ensure the interdependencies are properly understood, and support effective integration of the outputs of each workpackage, I will organise every 6 months **integration sprints** (see Gantt diagram). Integration sprints are one-week long integration retreat during which the whole WizUs team gather and work together to effectively implement and test on the robot the different components. In addition to providing regular 'check points' for the projects, they also set a stable schedule to deliver project components.

This methodology was adopted in a project the PI previously took part in (FP7 CHRIS project), and had proved at that time to be of great value to ensure project-wide cohesion and steady progress.

The three integration sprints taking place before the beginning of the experimental deployments (display as orange circles on the Gantt chart) are of particular importance, and will be extended to two weeks.



WP1: Framing robot-supported human-human interaction

The basic ambition of WizUs 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 WizUs Ethics Advisory Board, local ethics experts (including the head of the university ethics committee), and the WizUs 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 WizUs 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 [64] and [83]), 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 workpackages: WP2, WP3, WP4

The technical work programme of WizUs is spread over Workpackages WP2, WP3 and WP4. Figure 4.2 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.

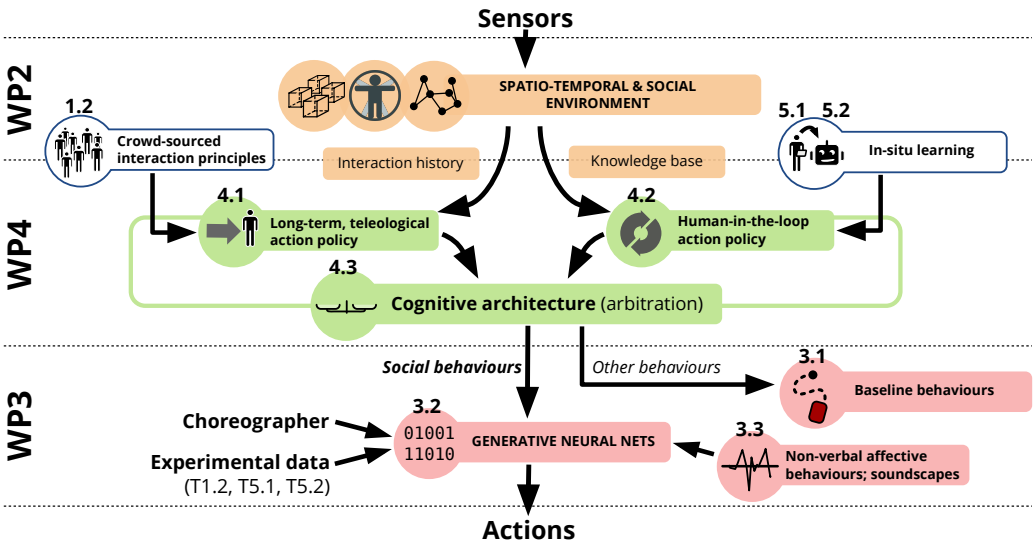


Figure 4.2: Overview of the AI engine implemented in WizUs.

WP2: 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).

T2.1 – Hybrid situation assessment and knowledge representation Knowledge representation

and grounding is a fundamental building block for cognitive architectures [49, 9]. This task builds on existing work on symbolic knowledge representation (eg [74] or my own work [46]) and my work on situation assessment [48] (that includes for instance object recognition and physics simulation [62]), to create a coherent system of representations for the cognitive architecture that extends the underworlds spatio-temporal representation tool [48] with symbolic and hybrid (like *conceptors* [35]) 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 [31] to build a multi-modal model of the humans in the robot's vicinity. I have recently introduced a dataset of social interaction [47] 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 [7]. 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 [73], including f-formations [52], sociograms (as done in [28] for instance), and inter-personal affordances [57]. This task builds on literature on on social dynamics analysis (eg [24, 36, 54]) 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 [25] (as I previously framed in [43, 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 [27] to investigate theory of mind with autistic children, as it offers an excellent experimental framework for social robotics [43] 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.

WP3: 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). WizUs 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 [55], Learning from Demonstration [11, 3], human-in-the-loop action policy learning [65, 64]. 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 [51, 71]; (2) interactive machine learning in high-dimensional input/output spaces, where I have shown with my students promising results for generating complex social behaviours [64, 83] that fully involve the end-users [84].

In [71], 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 [30]). 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.

WP4: 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 curiosity-driven robots [55] or motor babbling in infant-like robots [26], but only for relatively simple cognitive systems. 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 [64, 83], 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 [63].

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, 79, 37, 23, 39, 72, 75], as well as ones specifically developed for robotics: ACT-R/E [76], HAMMER [20], PEIS Ecology [60, 19], CRAM/KnowRob [9, 74], KeJia [16], POETICON++ [2], and my own, the LAAS Architecture for Social Interaction [49]): the overall purpose of the socio-cognitive architecture of WizUs 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 WizUs 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 [50] 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.

WP5: Experimental programme: long-term deployments in sensitive social spaces

WizUs 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 un-interrupted year. It will integrate the daily routines of the institutions, under supervised autonomy [63], 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.

TODO: 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 [59], 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 [80]), 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 [84] 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 ecosystem 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.

Ethics considerations and measures to ensure Responsible Research and Innovation

The WizUs 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 WizUs 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* [82])?
- 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 WizUs 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 WizUs 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 [69, 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 [53, 67]. 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 [66], have seen some additional work, but social robotics is still not equipped with operational guidelines, similar to the EU guidelines on trustworthy AI.

WizUs-specific measures

I have chosen to focus the first workpackage 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 WizUs 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 WizUs studies. It will be additionally supported by my extensive experience in seeking ethics approval for studies involving robots and vulnerable populations (in particular, children [45, 47, 64]), 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 [70] (and has been subsequently adopted by the UK Engineering and Physical Sciences Research Council [56]) using the acronym AREA: Anticipation, Reflection, Engagement and Action. The WizUs 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 [68] [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 expertise of the PI, and, as such, is low-risk. T1.1 can thus be described as a **medium-risk, high-gain** component of WizUs.

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 [48, 62, 46]. 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 [47, 61], and

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

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 [64, 83]. 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 [49]. 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 [49]. 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] S. Baron-Cohen, A. Leslie, and U. Frith. "Does the autistic child have a "theory of mind" ?" In: *Cognition* (1985).

[7] M. Bartlett, C. 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 Robotics and AI* (2019).

[8] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris, V. Enescu, A. Hiolle, I. Kruijff-Korbayova, R. Looije, M. Nalin, et al. "Long-term human-robot interaction with young users". In: *IEEE/ACM Human-Robot Interaction 2011 Conference (Robots with Children Workshop)*. 2011.

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

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.

[31] H. Gunes and B. Schüller. "Automatic Analysis of Social Emotions". In: Cambridge University Press, 2017, p. 213.

[32] 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.

[33] 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.

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

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

[36] 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.

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

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

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

[40] H. Lehmann, A. V. Sureshbabu, A. Parmiggiani, and G. Metta. "Head and Face Design for a New Humanoid Service Robot". In: *Social Robotics*. Ed. by A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He. Cham: Springer International Publishing, 2016, pp. 382–391. isbn: 978-3-319-47437-3.

[41] 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.

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

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

[44] 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.

[45] 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).

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, and G. Pioggia. "Autism and social robotics: A systematic review". In: *Autism Research* 9.2 (2016), pp. 165–183.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

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

[82] 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.

[83] 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.

[84] 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.

[85] 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.

B2.c Description of resources

Research team and PI commitment

Table 5.1 provides an overview of the time allocation per members of the team, over the course of the project.

	Y1	Y2	Y3	Y4	Y5	Total months
Séverin Lemaignan (PI)	0.6	0.6	0.6	0.6	0.6	36
Post-doc 1 (WP1)	1	1	1			36
Post-doc 2 (WP2)	1	1	1	1		48
Post-doc 3 (WP3)		1	1	1	1	48
Post-doc 4 (WP4, WP5)	1	1	1	1	1	60
PhD 1 (WP4, WP5)		1	1	1	0.5	42

Table 5.1: Full-time equivalent for the research team members

Team

PI Séverin Lemaignan will dedicate 60% (3 days/week) of his time to the project. This time will cover significant research time (about 2 days/week) as well as the supervision of the team and management of the project (1 day/week).

The rest of his time will be dedicate to other academic commitments within the Bristol Robotics Lab (including the on-going supervision of his other PhD students, supervision of MSc students, the supervision of the Human-Robot Interaction research group at BRL, lab-wide strategic engagement), as well as a small proportion of Master-level teaching in Human-Robot Interaction (about 5 days/term).

Each of the project work packages will have one lead researcher (post-doc); the duration of each of the post-docs' contracts roughly matches the duration of the corresponding work packages.

- WP1: I will appoint a post-doc (PD1) with a background in sociology of technology and science facilitation; the researcher will work for three year to frame the *robot-supported human-human interactions* paradigm, and lead the field work at the WeTheCurious science centre (to this end, the centre has committed to provide in-kind training in science communication to the researcher, enabling her/him to engage directly with the public);

- WP2: WP2 will be led by a post-doc (PD2) with a background in social signal processing and/or machine learning; the researcher will be appointed for 4 years; extensive collaboration with WP1's post-doc is expected to frame the social dynamics fostered by the robot;

- WP3: one post-doc (PD3, background in learning from demonstration and machine learning) will be in charge of developing the novel continuous robot behaviour generation method, and will be appointed for 4 years, starting on the second year;

- WP4: WP4 (the cognitive architecture) lays at the core of the project; the WP4 leader will be a senior post-doc in cognitive robotics (PD4), appointed for the whole 5 years to ensure continuity on this critical part; she/he will be responsible for the integration of the outputs of the other work packages; the same post-doc will also oversee (with the PI) the experimental work taking place in WP5.

The cost for the WP4 PhD student (PHD1) is *not requested*, as the host laboratory is part of the UK FARSCOPE Centre for Doctoral training, which will fund the student directly.

In addition, a small amount of budget is allocated to senior staff Dr. Dave Meckin (3 months FTE, support soundscape design, WP3.3) and Dr. Nigel Newbutt (4 months FTE, support the work in the SEN schools, WP5.1). I also have 3 month FTE of technician time allocated over the duration of the project to support specific technical developments on the robot.

Research equipment

I will purchase two IIT R1 robots (total €303,600; €379,500 incl. indirect costs) for the WizUs project. The R1 robot is a recently developed service robot from the Italian Institute of Technology (IIT). This purchase represents an additional cost with respect to the base €2M budget, for major equipment.

While the host institution (the Bristol Robotics Lab) will provide access to a range of social robots (some of them – PAL TiaGo and Softbank Pepper will be used for early prototyping), none of the currently available robots are fully suitable for the project. We provide a detailed comparison of the R1 robot features with respect to other social robots in section B2. Neither TiaGo nor Pepper (the two main alternatives) have non-verbal social features that are powerful enough to deliver the WizUs project. Critically, they both lack the abilities to show facial expressions or simulated gazing behaviours. Because the R1 robot features a programmable display in place of the head, we will have full freedom to create complex non-verbal facial expressions.

Besides, R1 has been designed from the ground-up to be used in care environments (in particular, hospital), and is made of materials that can easily be cleaned up/disinfected. This is of critical importance for the deployment in the hospital.

Two robots are necessary, to permit development on one platform while the other one is used in the field. In case of breakdown, the second robot will also be used as an emergency replacement for the first one, in order to ensure the continuity of the experiments. The two robots will be use exclusively by the project for the whole duration of the fellowship.

Travels

Travels and conference fees have been costed on the basis of one international conference per year and per person.

In addition, the budget include the costs of the four 2-days ethics workshops, that the three members of the ethics Advisory Board will be invited to join.

Subcontracting

The subcontracting amount covers: - the specific content creation and public communication costs, required to integrate the robot in the Bristol science centre WeTheCurious. - work with the choreographer from the RustySquid (http://www.rustysquid.org.uk/) company.

Open access

In line with the European requirements, all journal publications will made available under an Open Access license. On the basis of an average of 2 journal publications per annum, and an average processing fee of €1,200 per article, we request €12,000 to support Open Access costs. Note that conference publications do not always offer immediate open-access policies.

Other costs

The €5000 cost in section A.3 correspond to the project auditing.

Consumables include cloud computing resources, organisation of the ethics advisory board workshops, participant compensations.

Existing resources available to the researcher

The fellowship will take place at the Bristol Robotics Laboratory (BRL). The BRL is the largest co-located and most comprehensive advanced robotics research establishment in the UK. It is a joint venture between the University of the West of England and the University of Bristol. BRL has an international reputation as a leading research centre in advanced robotics research and has over 250 researchers working on a broad portfolio of topics, including HRI, collective robotics, neuro-inspired control, haptics, control systems, assistive robotics, soft robotics and biomedical systems. This multidisciplinary environment will directly benefit the project. BRL has many collaboration partnerships, both national and international, and is experienced in managing large multi-site projects. BRL has support from two embedded units specialising in business and enterprise, together with an incubator and successful track record of spin-outs.

The BRL also has a long track-record of designing and building new and original robots (from the BERT humanoid in the FP7 CHRIS project, to micro-robotics and surgical robots). WizUs will directly benefit of this expertise, which will ensure a feasible and realistic technical deployments of the WizUs robots. Dedicated technician time is allocated to this end.

The BRL also include a hardware incubator and is co-located with 70 start-ups and SMEs specialising in robotic hardware and mechatronics (Bristol's FutureSpace). This combination of excellent research and vast industry expertise on one site is unique in the UK, and is will play an instrumental role in providing opportunities beyond the project towards a strong pathway to impact, including further engagement with industrial partners and spin-off opportunities.

Other in-kind contributions

The Bristol science centre will provide in-kind training in science communication, as well as in-kind access to the centre facilities, for the duration of the study. The training (10 days in total) would have normally been billed £3,000 by the science centre.