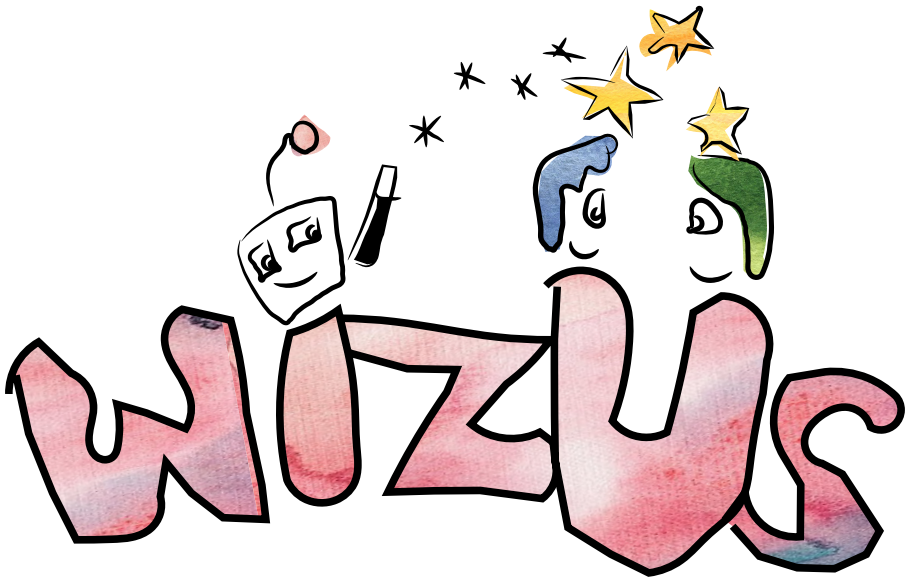
WizUs- Part B

January 29, 2020



Socially-Driven Robots to Support Human-Human Interactions

WizUs

- Principal Investigator: **Pr Séverin Lemaignan**
- Host institution: **University of the West of England, Bristol Robotics Laboratory**
- 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.

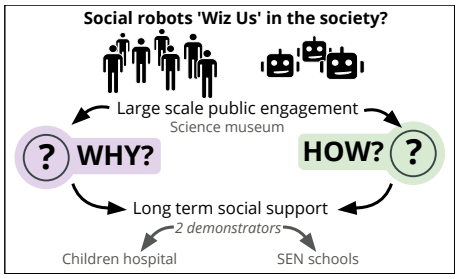
This breakthrough is made possible by a combination of novel methodologies and the principled integration of complex socio-cognitive capabilities:

- **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. Can we however ensure 'by design' that they will have a net social utility? WizUs defines and implements **a vision of AI and social robotics that aims at placing the human at the centre of these emerging technologies, to foster novel social dynamics that are accepted and perceived as beneficial by the society**. In other words, WizUs seeks to answer **why would we want to embed social robots in our society?**, and **how to do so?**, from a technology and Responsible AI perspective.

This research is ground-breaking: **WizUs main objective is, within 5 years, to design, implement and demonstrate the AI engine of a socially-intelligent robot, that will result in a robot that (1) has an effective social utility, and (2) will see long-term acceptance by its end-users**.

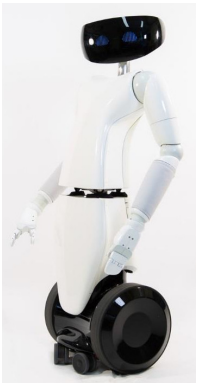


This objective is underpinned by two research hypotheses: (**H1**) for end-users to ascribe social utility and engage with the robot over long periods of time (months, years), the robot has to have its own long-term internal motivation to be socially helpful – what we call a *social telepology*. (**H2**) Additionally, long-term acceptance requires the genuine involvement of end-users at every step of the design process, so that they take *ownership* of the technology. I further hypothesise that human-in-the-loop machine learning is an effective way of generating ownership, by enabling the design of robot behaviours that genuinely originate from the end-users, and I therefore suggest that traditional human-in-the-loop machine learning could and should be extended into 'public-in-the-loop' machine learning, a new methodology to train a robot AI at scale, involving a large range of stakeholders and end-users, in different interaction situations.

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 museum *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 museum? 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 museum 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 8 children with long-term conditions stay for months, and engage with the children into playful social activities (eg 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 (IIT's R1, 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 (the focus of WP2) to represent the current social situation, 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 in addition create a novel non-verbal modality based on *soundscapes*: sound landscapes that the robot can modulate to influence the mood of the social environment (calm, excited, worried, etc.).

Finally, WizUs asserts and reinforces 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 to the building of this capacity in Europe**.

Overview of the WizUs work programme

WP1: Framing robot-supported human-human interaction

WP1 aims at establishing the conceptual amd 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 'public-in-the-loop' machine learning.

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 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 (T1.1, T5.1. T5.2) will both *build on* and *feed into* the framework developed in this task. In addition, four one-day workshops, spread over the duration of the project, will act as ethical milestones with key inputs from the WizUs ethics advisory board.

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*, *determinants* and *parameters* 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 WizUs in the Bristol' Science museum 'City Lab': for a whole year (Y2), one WizUs robot will be permanently based at the museum. With the help of a researcher, the museum 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 museum is 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 museum.

WP2: Real-world Social Situation Assessment

In WP2, the project addresses the key scientific and technical pre-requisites to effectively deliver WP4's 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); psychosocial 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 [35, 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 [34, 45] with recent advances in data-driven semantic labelling (for instance, using 4D convolution nets like MinkowskiNet [4]), and a symbolic knowledge base (like my own ontology-based one [29]) in order to create a coherent system of representations for the cognitive architecture of the robot.

T2.2 – Social dynamics This task focuses on the processing and modelling of social signals, extending existing techniques, both model-based (eg [**gunes2017automatic**, **others**, 27]) 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 [31], as well as from the project experiments (T1.1, 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, including f-formations [39], sociograms (as done in [11] for instance), and inter-personal affordances [43]. This task builds on literature on on social dynamics analysis (eg [18, 40]) 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 [10], that I introduce in [24]) 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 behaviour, 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 HRI/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 [5]). 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 [38, 50]; (2) interactive machine learning in high-dimensional input/output spaces, where I have shown with my students promising results for generating complex social behaviours [48, 56] that fully involve the end-users [58]. 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 need treatments (eg [13]). 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.

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 [14], 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 [55], 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.

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 simple cognitive systems [42, 9]). 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 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 [46, 48, 56], this task implements the mechanics to allow human end-users to progressively teach the robot a social policy to become a effective social helper. 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 [35], 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 [36] that enables the re-assessment of a policy (here, our intrinsic motivation) based on previous experience (here, the human-taught policy).

WP5: Evidence-based research: demonstrable usefulness of social robots in real-world, complex scenarios

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 addition to the museum study, T1.2), 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 positive 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 [58] to design the role of the robot with the different stakeholders (nurses, doctors, parents, children), we 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

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.

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

Psycho-social underpinnings of HRI	
human factors	anthropomorphism [25], cognitive correlates [26], social influence [57]
trust, engagement, social presence	[8, 32, 7, 16]
theory of mind	perspective taking [44, 54], social mutual modelling [24, 6]
Socio-cognitive architectures	
architecture design	[35, 3, 23, 22, 37]
knowledge representation	ontologies [29, 30]
spatio-temporal modelling	object detection [51], physics-aware situation assessment [34, 45]
Social signal processing	
non-verbal behaviours	attention [27], child-child dataset [31], internal state decoding [1]
verbal interactions	speech recognition [20], dialogue grounding [33]
Behaviour generation	
social behaviours	[21], verbal interactions [52, 53], physical interactions [12]
interactive reinforcement learning	[49, 46, 48]
Fieldwork in HRI	in classrooms [15, 28, 17, 2, 19, 47], at home [41]

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 moved from being a regional leader in HRI, to having established myself as a world-leader in the emerging field of socially-driven, responsible autonomous robots, building up the European capacity in this critical field for our digital future.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] I. García-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.

[12] M. Gharbi, S. Lemaignan, J. Mainprice, and R. Alami. "Natural Interaction for Object Hand-Over". In: 2013.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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).

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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).

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

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

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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

Pr. Séverin Lemaignan

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

EDUCATION

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

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

PREVIOUS POSITIONS

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

FELLOWSHIPS AND AWARDS

2019 **UWE Vice Chancellor Accelerator Fellowship**

2015 – 2017 **EU Marie Skłodowska-Curie Individual Fellowship** Theory of Mind and social robotics
Plymouth University, UK

HRI'2017 Best Paper award

HRI'2016 Best Paper award

AAAI'2015 Best Video award in Artificial Intelligence

HRI'2014 Best Late Breaking Report award

2012 **Best PhD in Robotics 2012** award, CNRS, France

2012 PhD with **High Distinction** ("Summa Cum Laude"), TU Munich

Ro-Man'2010 Best paper award

SUPERVISION OF GRADUATE STUDENTS AND POSTDOCTORAL FELLOWS

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

TEACHING ACTIVITIES

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

ORGANISATION OF SCIENTIFIC MEETINGS

2020 **ACM/IEEE Human-Robot Interaction conference**, 700+ participants, local chair, Cambridge, UK
2017 **ACM/IEEE Human-Robot Interaction conference**, 400+ participants, alt.HRI chair, Vienna, AT
2016 **2nd Intl. workshop on Cognitive Architecture for Social HRI**, 45 participants, programme chair, Christchurch, NZ
2014 **Intl. workshop on Simulation for HRI**, 35 participants, programme chair, Bielefeld, DE
2012 **Intl. workshop on MORSE and its applications**, 30 participants, programme chair, Toulouse, FR
2009 **Cognitive Sciences' Young Researchers Conference**, 150 participants, steering committee, Toulouse, FR

INSTITUTIONAL RESPONSIBILITIES

2019 – Associate Professor, Faculty of Technology and Environment, UWE, UK
2019 – Head of the Outreach cluster, Faculty of Technology and Environment, UWE, UK
2019 PhD defense committee, University of Bielefeld, DE
2019 PhD defense committee, University of Örebro, SE
2018 – HRI module co-lead, MSc level, University of the West of England, UK
2017 – 2018 Module leader, Robotics fundamentals (undergraduate level), University of Plymouth, UK

EDITORIAL ACTIVITIES

2018 – Editorial board of *Frontiers in AI and Robotics*
2015 – Member of the IEEE/ACM HRI Programme Committee
2019 – Member of the Robotics, Science and System (RSS) Programme Committee
2017 – 2019 Member of the IEEE IROS Programme Committee
2017 – 2018 Member of the IJCAI Programme Committee
2017 – 2018 Member of the HAI Programme Committee

Appendix: Current research grants and any on-going applications related to the proposal

Current Grants

Project Title	Funding source	Amount	Period	Role of the PI	Relation to current ERC proposal
CAPRI	InnovateUK (UK)	€4 840 508	2017 – 2020	Co-I for BRL; driverless car simulation for safety verification	Dev. and verification of trustworthy autonomous systems
ROBOPILOT	InnovateUK (UK)	€7 986 981	2018 – 2020	Co-I for BRL; driverless car simulation for safety verification	Dev. and verification of trustworthy autonomous systems
CAV Forth	InnovateUK (UK)	€5 093 327	2019 – 2021	Co-I for BRL; supervising the safety case and simulation-based verification	Dev. and verification of trustworthy autonomous systems
RoboClass	UWE (UK)	€5 854	2019 – 2020	PI; project supervision and robot development	Classroom deployment of a social robot

On-going and submitted grant applications

Project Title	Funding source	Amount	Period	Role of the PI	Relation to current ERC proposal
RoboPets	Amazon (US)	€10 942	2020 – 2020	PI; project supervision and lead researcher	Learning and generation of continuous, congruent social behaviours
ROBUST	EPSRC (UK)	€761 124	2020 – 2022	PI; project supervision and architecture implementation	Dev. of a redundant cognitive robot architecture for HRI
HEROS	H2020 (EU)	€7M	2021 – 2024	Co-I for BRL; research lead on cognitive architecture	Dev. of a redundant cognitive robot architecture for HRI
Robots4SEN	UWE (UK)	€29 274	2020 – 2021	PI; project supervision and robot development	Pilot deployment of a social robot in a SEN school

B1.c Early achievements track-record

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

Soon after my PhD, I initiated and successfully led for 2 years the HRI group within the AI for Learning CHILI Lab at EPFL (Switzerland), supervising in total 10 students, and creating in a short timeframe an internationally visible centre of excellence in educational robotics. While my original training was in **symbolic cognition & AI, and software engineering for autonomous robotics**, my postdoctoral stay at the highly cross-disciplinary CHILI Lab gave me the opportunity to develop my expertise in **experimental sciences, socio-psychology and education sciences**.

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

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

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

This expertise is recognised internationally: I have a substantial track record of academic outputs (since 2008, I have authored or co-authored **75+ peer-reviewed publications** in international journals and conferences, leading to **2200+ citations**, h-index of 24, i10-index of 39 (source: Google Scholar).

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

- invited to **high-profile editorial roles**: Programme Committee member of the HRI conference since 2015; editor of Frontiers In Robotics and AI journal; editor or Programme Committee member of several leading conferences in AI and Robotics (IROS, IJCAI, HAI, AAMAS);
- invited member of the UK EPSRC Associate Peer Review College;
- numerous **invited talks** at national and international symposiums and events (9 invited talks since Jan. 2018, including **keynotes** at the UK Robotics and Autonomous Systems 2019 conference, and at the 2018 AAAI Fall Symposium);
- local **organiser for the high-profile, international HRI2020 conference**;

Research dissemination

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

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

give examples of coverage

Tech transfer; CAV; SABRE; patent US20190016213A1

Selected outputs (reverse chronological order)

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

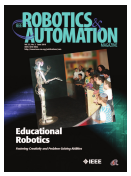
A novel representation technique to efficiently represent multiple parallel states of the world, including imaginary ones. This ability is critical to represent spatio-temporal predictions, and to create models of other agents' representations. **[principal investigator]**

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]

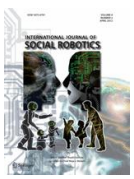
Lemaignan, S., Warnier, M., Sisbot, E.A., Clodic, A., Alami, R.
Artificial Cognition for Social Human-Robot Interaction: An Implementation
Artificial Intelligence 2017

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 Module for Cognitive Architectures in Robotics
IEEE IROS 2010

One of the very first knowledge base designed and integrated in service robots. Pioneering work which played a key role in understanding how intelligent robot can represent their knowledge to facilitate communication with humans.
[principal investigator; 155 citations]

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

TARGET PAGE COUNT: 4 pages: state of art + vision; 2 pages: methodology overview; 4 pages: WPs; 3 pages: ethics + risks; 2 pages: resources

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

THIS SECTION IS STILL WORK-IN-PROGRESS

rewrite section, mostly borrowed from [87]

Social robotics is a disruptive field, with a profound impact on society and economy [91]. 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 [21]. The impact of AI applications and manufacturing robots on economies and labor markets is already tangible. Yet, this is not the case for social robots despite the fact that this technology is expected to have a significant impact on different application areas such as care for the elderly, customer service, education, child development, and autonomous vehicles [5].

As a matter of fact, in the past years promising companies are facing crises. Some of them are being bought by multinationals, such as Aldebaran, Boston Dynamics and Scharf, acquired by Softbank Group, while others shutdown (e.g., Willow Garage, Anki, Jibo). Developing and selling robots is challenging and require market knowledge that companies born as a spin-off from Universities may not have. This can be the case of Rethink Robotics founded by Rodney Brooks and Jibo co-founded by Cynthia Breazeal, both of them from the Massachusetts Institute of Technology. After releasing the collaborative robot Baxter and its counterpart Sawyer, Rethink Robotics has been acquired by HAHN Group, a German automation specialist. Jibo stopped operating nearly one year after it first came to the market. The same end has been encountered by Anki, a robotics and artificial intelligence startup founded by three graduated students from Carnegie Mellon University.

The failures of companies like Jibo, Kuri, Willow Garage and Anki illustrate the complexity of bringing social robot into real-world applications with sustained engagement. The scifi-inspired general-purpose robot companion has not realised yet, and successful robots are indeed very much task-specific: autonomous cleaning devices (e.g., iRobot Corporation, Samsung Electronics, Neato Robotics, LG Electronics), surgical robots (e.g., Intuitive Surgical, MAKO Surgical Corporation), drones (e.g., Parrot SA, 3D Robotics, D-Jing Innovations Science and Technology) and toys (e.g., Hasbro, WowWee Group Limited).

Examples of successful commercial social robots are robotic pets developed in largest public research organization and companies, such as Paro Therapeutic Robot (AIST), PLEO (Innvo Labs established by Jetta Corporation), AIBO (Sony Corporation), Keepon (BeatBots), and iCat (Philips). Other examples can be found in the context of smart home assistants; Mykie robot is a side project of BOSCH connected to an IoT ecosystem. The robot is designed to support users while they cook providing assistance for recipes and stock market prices. Mykie robot is activated and controlled by Amazon's Alexa, conversational virtual agents that nowadays dominates the market of smart homes devices [67, 72].

Having social capabilities embedded in robotic systems provide an additional emotional grip and more fluid interaction with humans, but is not enough for justifying the needs of social robots over long periods [46, 6]. The human-robot interaction should be driven by clear scopes, and the robot should be able to adapt to diverse and unstructured environments. However, notwithstanding the improvements in computer vision and machine learning, very few fully-developed intelligent autonomous systems capable of learning from the real-world and successfully interact with humans are currently available to consumers [28]. Thus, all these companies are trying to adapt their business principles and products for surviving against competitors breakthrough technologies and new customer demands. The paradoxes encountered by the research and industry involve three main actors: final users, organizations (companies) and researchers.

[40]

One example of a small, rugged robot designed for intensive use in school environments is Cellulo[1].

Looking specifically at human-sized mobile manipulators with advanced social features, the choices of robotic platforms are limited. Table 4.1 compares the two leading mobile social robots available on the market today, along with the new R1 platform developed by the Italian Institute of Technology. While not yet on the market,

[1] 63.

Table 4.1: 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 advanced social interaction capabilities.

	PAL TiaGo	Softbank Pepper	**IIT R1**
Social features	Poor (non-expressive head)	Medium (expressive, yet fixed, face; limited gaze; non-threatening appearance)	Good (expressive face [45]; artificial skin for touch-based interactions; non-threatening appearance)
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)

the IIT has offered early access to the platform for this project. Note that the Fetch Mobile Manipulator has been omitted, but is functionally very similar to the PAL TiaGo.

Sustaining long-term interactions In his analyse of why many commercial projects around social robotics failed, Hoffman cites the lack of long-term acceptance as one major issue [40]. He offers as key explanation "the inability of the robots to escape the single turn structure of an interaction", which also tightly connects to the issue of the repetitiveness of the robots' behaviours.

[28]

Social robotics and vulnerable children

Using activity switching to support long term engagement with diabetic children [25] Long-term engagement in a paediatric ward [14]

Robotics and autism [65]

The false belief experiment that we have mentioned above, was proposed by Baron-Cohen in the frame of his research on autistic spectrum disorders (he shows that autistic children seem to actually lack a theory of mind and suggests this as the primary cause of their social impairments), and Frith and Happé further note in [35] that this specific deficit of autism has led to a large amount of research which proved, in turn, highly beneficial to the study of the development of theory of mind in general. They reference in [35] eight such tasks (Table 5.1), identified during the study of social cognition by autistic children. Each of them is proposed in two versions: one does not require mentalizing, while the other does require it. One of these tasks, for example, required children to distinguish emotions, namely happy/sad faces on one hand (*situation−based* emotion), and surprised faces on the other (*belief−based* emotion) [11]. Another task, based on the *penny−hiding game*, contrasts the two conditions in terms of *object occlusion* vs. *information occlusion* [8] (we detail it hereafter). These tasks prototypically illustrate social meta-cognition: one need to represent and reflect on someone else representations (and not only perceptions), and they are not addressed by today's research on social robots.

Long-term social engagement is beyond-state-of-the-art

As a researcher who has been working for the last 12 years in the field of human-robot interaction (and child-robot interaction in particular), I have been a direct witness – by being one of the architects – of the crossing of a critical milestone: the emergence of **long-term social interactions** between robots and humans [43, 46], with a number of studies involving social robots deployed in real-world settings (schools [90, 50], care centres [39]) over relatively long periods of time (up to 2 or 3 months at a time). Even though these robots are rarely fully autonomous, they do already show high levels of autonomy [74], with full autonomy in sight [39] for simple tasks.

However, the question of the long-term *social engagement* is open.**finish that**

Some guidelines:

[87] [40]

[85]

[6]

In a field more closely related to child-robot interaction, the large review of research in robotics for education published by Belpaeme et al. in 2018 [17] points to the major shortcomings that prevent further development of social robots in educative settings: the need for a correct interpretation of the social environment; the difficulty of action selection; the difficulty of pacing generated behaviours.

WizUs objectives: Responsible robots for long-term social engagement

THIS SECTION IS STILL WORK-IN-PROGRESS

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 generation 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.

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 requires the genuine involvement of end-users at every step of the design process, so that they take *ownership* of the technology. We further hypothesise that human-in-the-loop machine learning is an effective way of generating ownership, by enabling the design of robot behaviours that genuinely originate from the end-users, and we therefore suggest that traditional human-in-the-loop machine learning could and should be extended into 'public-in-the-loop' machine learning, a new methodology to train a robot AI at scale, involving a large range of stakeholders and end-users, in different interaction situations.

Key scientific challenges and research questions

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, WP4 and WP3), 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 questions:

- **R1** [conceptual framing]: what are the basic principles of responsible social interactions, that must form the foundations of a socially useful robot, accepted and used in the long run? What should motivate the robot to step in and attempt to help? What are the determinants and parameters of a social intervention, performed by a socially-driven robot, to support positive human-human social interactions? How to balance social utility and social responsibility?
- **R2** [implementation level]: how will these principles be integrated into a principled, socially-driven teleological architecture for autonomous robots? How this should be combined with bottom-up action policies, designed and learnt from the end-users? How can we ensure 'by design' that the resulting AI will generate useful yet responsible, trustworthy, human-centered robot behaviour?
- **R3** [technology level]: where are the technological gaps in artificial social modeling and cognition, that prevent the actual realisation of a robot capable of effective social support, sustained over long period of time? How can we fill them?
- **R4** [experimental level]: can we demonstrate in complex, real world conditions, the effectiveness and usefulness of the robot-supported human-human interaction paradigm? Can we do so by involving the end-users at every stage of the design, implementation and testing cycle?

WizUs is also a highly technical project, who aims at significantly pushing the state-of-the-art in autonomous social robotics. Indeed, in WizUs, we will **implement the AI required for robots to effectively support human social interactions**. In that sense, this research is also ground-breaking in regards to its technical objectives. In WizUs, robots will be able to understand complex social dynamics, and generate appropriate social responses, in a fully autonomous way. Extending the current line of research of the PI, we will identify, implement, and integrate the a broad range of cognitive functions into a principled, socially-driven, and trustworthy socio-cognitive architecture for robots. This is the second major expected scientific outcome of WizUs.

Interdisciplinary nature of the research programme

Grounded in both the psycho-social literature of human cognition, and the latest technological advances in human-robot interaction, the project delivers major conceptual, technical and experimental contributions to the field of AI, with a particular focus on ethics & safeguarding mechanisms, in order to build 'by design' a trustworthy AI system.

WizUs delivers this programme by building on a range of multidisciplinary methods, including sociological investigation, novel interactive machine learning techniques, and a pervasive approach to co-design that puts the end-user needs at the centre of the design process. It paves the way for a better understanding of the societal challenges raised by the rapid development of AI and robotics, and, critically, opens a **unique window into what positive role social robots could play in our future societies**.

The interdiscplinary nature of the project is also reflected in the complementary research profiles that will be recruited on the project. **finish that**

Impact: artificial embodied social cognition & trustworthy social robots

THIS SECTION IS STILL WORK-IN-PROGRESS

WizUs will deliver new and fundamental knowledge to the fields of robotics, computer science and psychology, in addition to improving trans-disciplinary understanding between these disciplines. More specifically, the project will contribute to a better understanding of the following research areas: human-robot interaction; human-machine interaction; human error making and handling in assembly tasks; theory of mind and its transfer to cognitive robots; natural language processing; explainability and language generation; machine vision and human activity detection; and action planning under uncertainty. New interaction paradigms will be developed for handling error situations where machines interact with non-expert humans. The integration of perception and sensing into cognitive robot architectures will be critically reviewed and extended. Novel, empirically informed methods to transfer findings from human-human interaction studies to human-machine interactions will be developed. Furthermore, the national and international psychology, robotics and computer science research communities will benefit from the project results. We will publish WizUs results in interdisciplinary and high-profile discipline-specific journals (eg Science Robotics; Frontiers in AI and Robotics; Transaction in Human-Robot Interaction and conferences (AAAI, HRI, RSS), and organise WizUs-themed workshops.**details**

Academically, the WizUs project represents a timely combination of very recent advances in supervised machine learning for social robot behaviour with a creative and interdisciplinary approach to the design and automation of social robot behaviour. We therefore expect to publish results in high-class scientific journals and conferences.

The dataset of social behaviours and social signals we will create and distribute represents a one-in-a-kind resource for the human robot interaction community, and the human data collection will be transferable to research in other domains such as human-computer interaction.

As WizUs will be deployed in a living lab environment, there is significant scope for public outreach/engagement and media coverage, which we will work with the BRL's media manager to maximise.

WizUs aims at building unique European capacity to assert leadership in this domain, and, beyond the specific deliverables of this 5-years project, establishing the PI as a world-leader in goal-driven, socially-responsible robotics.

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** trans-

poses 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).

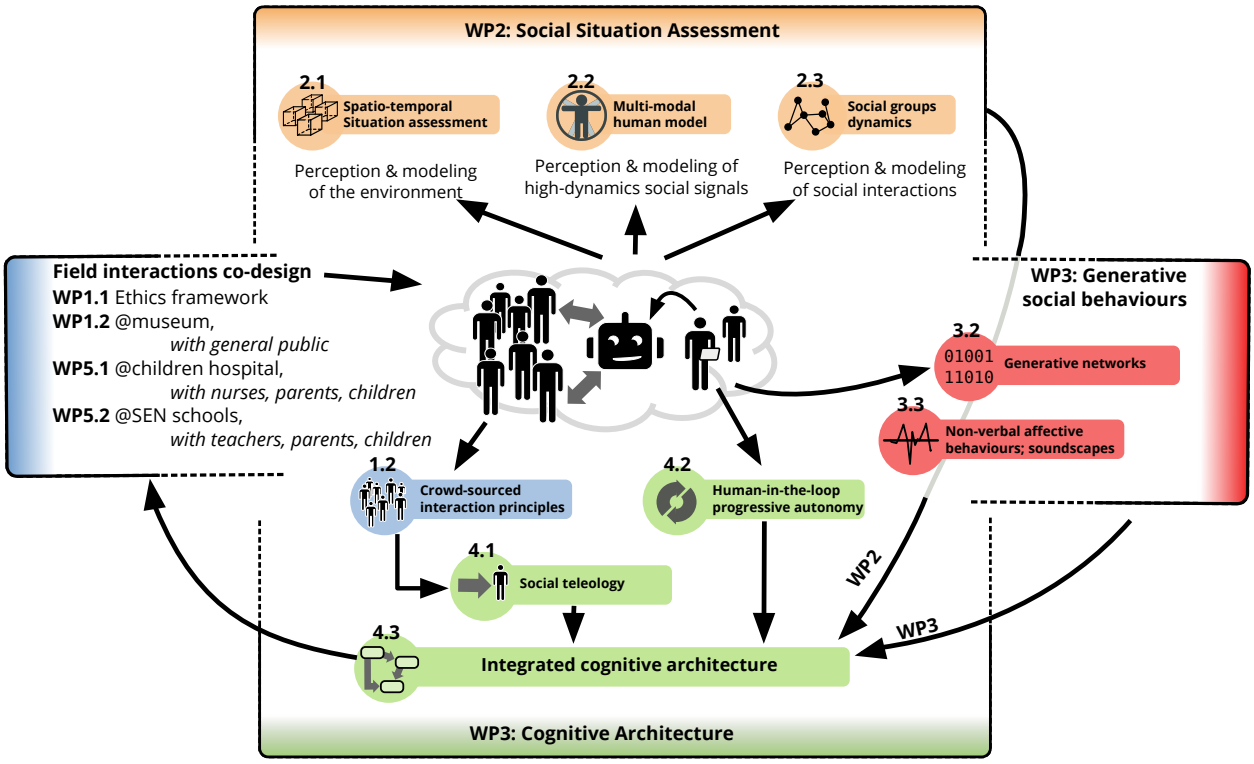


Figure 5.1: Overview of the workpackages and tasks, and tasks inter-relations.

More specifically, Figure 5.1 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 museum (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.

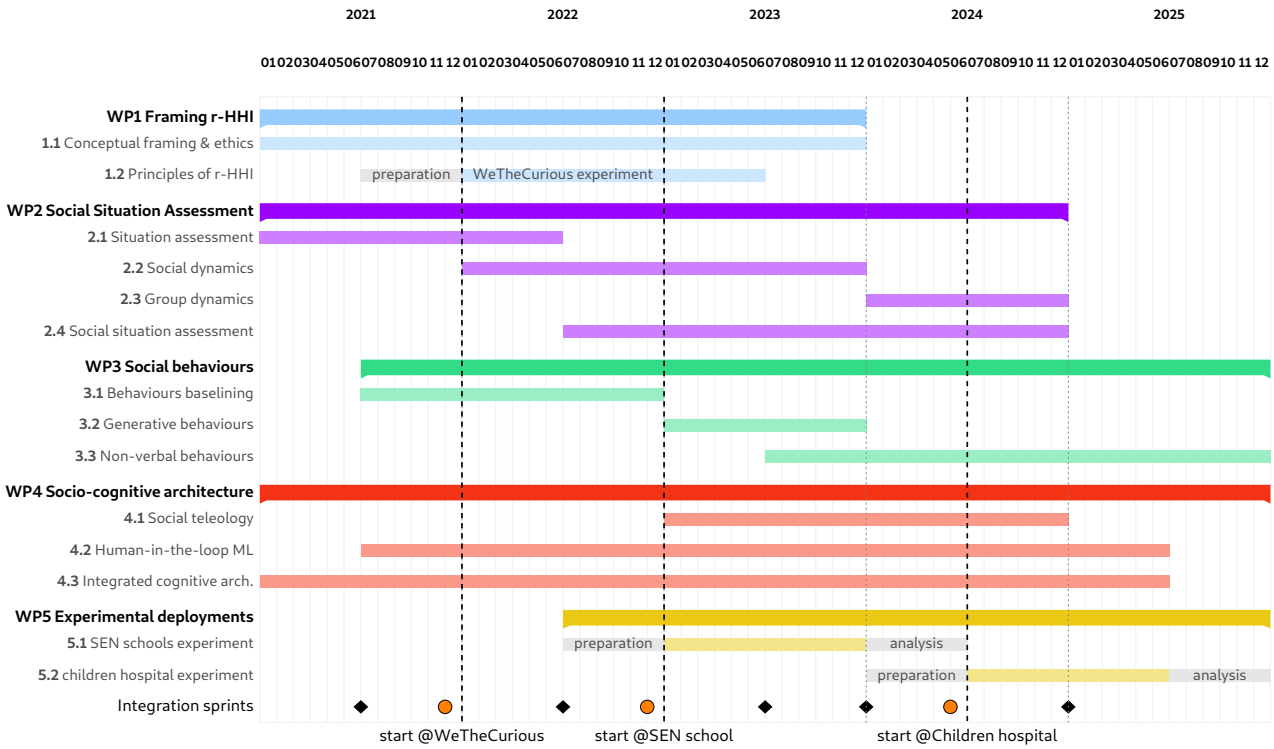
Integration sprints

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 in orange on the Gantt chart) are of particular importance, and will be extended to two weeks.



WP1-WP5 SECTIONS ARE STILL WORK-IN-PROGRESS

WP1: Framing robot-supported human-human interaction

The basic ambition of WizUs is to create a conceptual framework around social robots in the human society, by re-framing the traditionally accepted idea of *human–robot interaction* into the human-centered idea of *robot–supported human–human interactions* (r-HHI): the robot is considered from the perspective of how it can *sup–port* humans, and in particular, support stronger, positive interactions between humans.

T1.1 – Conceptual framing of r-HHI The first task in WP1 is to research and define such a framework that will provide the (currently missing) conceptual frame around questions like: what role for social robots? where to set the boundaries of artificial social interactions? what does 'ethical-by-design', 'responsible-by-design' might mean in the context of social human-robot interactions?

In order to anchor T1.2 into the reality and complexity of human social interactions, and to also involve the civil society in this framing process, the task will embed WizUs into the 'City lab' experiment, conducted by Bristol's science museum 'WeTheCurious'. WeTheCurious has developed a new form of public engagement called 'City Lab', the first in the UK that sees the visitors engaging in the actual production of science. We will integrate WizUs in the City Lab to co-design and co-produce robot-supported social interactions with the general public. For an initial period of one year (Y2-Y3), one WizUs robot will be permanently based at the museum. Participants (children and adults) will be guided, with the help of museum staff and a dedicated interface, into teleoperating the robots to make them good 'social helpers'. 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 large-scale participatory design with future end-users of social robots.

Specific resources I have an on-going collaboration with WeTheCurious, and preliminary meetings were held to discuss specific requirements for the WizUs project. The museum is committed to the project, and will include WizUs in its official programme of activities.

T1.2 – Determinants and principles of robot-supported social interactions The conceptual framework identified in T1.1 is translated into a set of *interaction design principles*, *determinants* and *parameters* that will to-

gether form a set of requirements and objectives for the socio-cognitive capabilities and architecture developed in WP2 and WP4.

WP2: Real-world Social Situation Assessment

In WP2, the project addresses the key scientific and technical pre-requisites to effectively deliver WP4's architecture: the scientific understanding and formalisation of the *social fabric* in which the robot is embedded, in its full complexity: 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). While several of these capabilities have been previously investigated is isolation [48, 33, 53, 32, 68, 89, 47, 29, 94], this WP will deliver the first complete and integrated model of artificial cognition that account for social interactions in their full extend, significantly extending the state-of-the-art [56, 13].

T2.1 – Hybrid situation assessment and knowledge representation Knowledge representation and grounding is a fundamental building block for cognitive architectures [56, 15]. This task builds on existing state-of-the-art in knowledge representation and situation assessment (eg [**citeneeded**]) and creates a coherent system of representations for the cognitive architecture that extends the underworlds spatio-temporal representation tool developped by the PI [55, 70] with knowledge representation capabilities, using both established symbolic techniques (like ontologies and first-order logic [51, 83]), and hybrid symbolic/sub-symbolic modelling (using Jaeger's conceptors [41]) as a new route to overcome the symbolic grounding problem [38].

T2.2 – Social dynamics This task focuses on the processing and modelling of social signals, extending existing techniques, both model-based (eg [**others**, 49]) and machine-learning based (eg [**chetouani**, **others**]) This task goes beyond the state-of-the-art by looking specifically at resolving highly dynamical signals (like gaze saccades and micro facial expressions). While playing a fundamental role in social interactions [**citeneeded**], they are currently not investigated in social robotics – even though the technology (high speed cameras and embedded GPUs) to achieve real-time classification of such cues is available.

T2.3 – Interaction and group dynamics Building on T2.2, T2.3 investigates the automatic understanding and modelling of group-level social interactions, like inter-personal affordances [64]. It includes spatial determinants (proxemics; group-level attention tracking); psycho-social determinants (social roles and hierarchies; social groups) and dynamics (effects of novelty; dynamics of anthropomorphism and mental ascriptions; group dynamics).

T2.4 – Social situation assessment 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 we 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 theory of mind (as I previously framed in [47, 29]).

A focused experimental programme accompanies T2.4, to demonstrate (in relative isolation) the resulting socio-cognitive capabilties. In particular, the protocols identified by Frith and Happé [35] to investigate theory of mind with autistic children offers an excellent experimental framework for social robotics [47] for this work.

Experimental protocols in research on autistic spectrum disorders are often striking by their apparent straightforwardness because of the careful choice of interaction modalities: since autistic children frequently exhibit impairments beyond social ones (such as motor or linguistic ones), the experiments must be designed such that they require only basic cognitive skills beyond the social abilities that are tested. The Sally and Anne task, for instance, requires the observing child to be able to visually follow the marble, to remember the true location of the marble, to understand simple questions ("Where will Sally look for her marble?" in Baron-Cohen's protocol [7]) and eventually to give an answer, either verbally or with a gesture – the two first points being actually explicitly checked through questions: "Where is the marble really?" (reality control question) and "Where was the marble in the beginning?" (memory control question).

Likewise, current social robots have limited cognitive skills (no fast yet fine motor skills, limited speech production and understanding, limited scene segmentation and object recognition capabilities, etc.) and such tasks that effectively test a single cognitive skill (in this case, mentalizing) in near isolation are of high relevance for experimental social robotics.

Frith and Happé's list (Table 5.1) is in that regard especially interesting in that it mirrors pairs of task (ones which do not require mentalizing with similar ones which do require mentalizing), thus providing control tasks. *Object occlusion* vs. *Information occlusion* is one example of a (pair of) task(s) which evidence representation-level perspective taking through *adaptive deception*: during a simple game, the experimenter adapts its strategy

No mentalizing required	Mentalizing required
Ordering behavioural pictures	Ordering mentalistic pictures [10]
Understanding see	Understanding know [66]
Protoimperative pointing	Protodeclarative pointing [9]
Sabotage	Deception [78]
False photographs	False beliefs [57]
Recognizing happiness and sadness	Recognizing surprise [11]
Object occlusion	Information occlusion [8]
Literal expression	Metaphorical expression [37]

Table 5.1: Tasks requiring or not mentalizing to pass, listed by Frith and Happé in [35]

(deceptive/non-deceptive behaviour) to the representation skills of its child opponent. The experimental setting is derived from the penny-hiding game protocol originally proposed by Oswald and Ollendick [60] and replicated and extended by Baron-Cohen in [8], who describes it as a two-person game in which the subject is actively involved, either as a guesser or as a hider. The hider hides the penny in one hand or the other, and then invites a guess. The game is repeated several time before switching the roles. Baron-Cohen proposes a specific index to rate the level of the players based on the idea of *information occlusion*: minimally, the hider must ensure *object occlusion* (the penny must not become visible to the guesser), while good hiders, with representation-level perspective taking skills, develop strategies (like random hand switching or deictic hints at the wrong hand) to prevent the guesser to find the penny (*information occlusion*). One could imagine a similar protocol adapted to robotics: the robot would play the role of the experimenter, adapting on-line its behaviour to what it understands of the perspective taking capabilities of the children, and would consequently require *second−order*, *representation−level* perspective taking from the robot.

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

Using activity switching to support long term engagement with diabetic children [25]

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 [61], Learning from Demonstration [18, 3], human-in-the-loop action policy learning [75, 74]. This baseline will enable early in-situ experimental deployments (WP5), while also provide a comparison point for T3.2.

T3.2: Machine learning for continuous motion generation

WizUs aims at significantly advancing the state of the art in this regard, by combining two existing techniques: (1) data-driven, continuous approach to behaviour generation inspired by Learning from Demonstration; (2) interactive machine learning in high-dimensional input/output spaces [**senft2020woz**], where I have shown with my students promising results for generating complex social behaviours [74, 93] that fully involve the end-users [95]. By combining the two, I target a breakthrough in robots' social behaviours generation: the generation of non-repetitive, socially congruent and transparent social behaviours (including gestures and gazes).

Designing behaviours that enable sustained, long-term engagement in a social human-robot interaction is essentially an open research question. Three main approaches to social behaviours generation exist today: *user−*

induced, where the end-user interacts with the robot and ascribes (knowingly or not) complex behaviours to the machine, while in reality the robot's behaviours are simple and non-goal oriented (eg generating a noise or a small movement when being touched). This has been used to great effect in therapy robots, for instance (eg Paro). *Off-the-shelf behaviours*, where the robot relies on a set library of behaviours (that might be individually relatively complex). The approach can elicit a strong initial social response from the user, but this social response tends to vanish rapidly once the 'tricks' of the robot have been all discovered and become repetitive. Besides, as the robot does not typically maintain a long-term socio-cognitive plan of the interaction, the behaviours are typically perceived as fun, yet pointless. This is often observed in toy-like robots (eg Vector, Dot & Dash). Finally, many social robots avoid altogether the problem of generating behaviours by simply offering to the end-user control over *low-level behaviours* (eg, control of the joints of the robot). This means that, even when the robot has relatively powerful social perception capabilities (like recognising people and voice), no real social behaviours is generated.

None of these three approaches are satisfactory, and indeed, no approach to date has been able to engage human users in long-term, sustained interactions.

At a time where companion robots are coming to the market, one important question remains fully open: how to design robot behaviours that foster lasting engagement? A vast body of academic literature identifies that robots evoke an initial phase of high user engagement (the *novelty* phase) that vanishes as the user realises that the robot is actually quite predictable and repetitive. The *agency* initially ascribed by the user to the robot quickly fades [48], leading to critical user disengagement from the technology.

The (often limited) library of behaviours available to the robot is often cited as a key factor in causing this issue. However, another, more profound issue affecting long term engagement with robot companions is the question of *purpose.* Without clear *purpose*, social robot companions can lack *usefulness*. Indeed, robot *companions* might not have explicit goals that would dictate or motivate their behaviours: they aim at providing a social presence, a social comfort, as cats or dogs would do, without necessarily being goal-oriented.

Recent attempts – and failures – to convert social robotics research into commercial platforms (Jibo, Kuri and most recently Anki's Cozmo and Vector robots) reflect exactly this, with reasons for their failure typically citing an under-delivery of the user experience they promised, and/or the lack of a 'real need' to justify their price point. The WizUs project addresses these two key issues by:

1. Taking inspiration from human-pet relationships which also have no explicit *purpose* beyond their potential for enjoyable, *affective* interactions;
2. Working with creative professionals who excel at storytelling and emotional engagement to overcome the problems in sustaining engagement, as proposed by Hoffman [40]
3. Blending these two sources of inspiration using a radically novel combination of immersive teleoperation and machine learning.

The project is *not* about replicating a pet's behaviour per se. It is instead about identifying, modeling and automatically generating the social behaviours required to recreate pet-like social dynamics between robots and humans, drawing inspiration from ethology (Stanton, Sullivan, and Fazio 2015). Using animal behaviours to inform the design of robots is not new, the most remarkable example being the Sony AIBO robot dog, whose behaviours were directly designed around those of actual dogs (Arkin et al. 2003). However, to go beyond the repetitive interactions associated with such robots, we propose to employ a creative professional to actively participate in design and automation of WizUs behaviour. The concept of using creative professionals to 'teach' social robot behaviour is not new either (Knight and Gray 2012), however it is only recent advances in human-in-the-loop, online machine learning that make this type of real-time 'social training' a feasible approach to generating and automating engaging social behaviours (Senft et al. 2019).

Our project has the following goals, addressed by the workplan presented below:

1. assemble a non-anthropomorphic social robot that can autonomously navigate in a complex and living lab environment, taking inspiration from ethology to inspire the robot's behaviour;
2. develop an immersive teleoperation system, enabling a creative professional to 'take control' of the robot (i.e. puppet the robot) in a completely intuitive way (using whole body motion tracking);
3. record (and make publicly available) a large dataset of social behaviours (created through immersive teleoperation) that foster long-term social and affective engagement. The dataset will also include the social *signals* implicitly used by the puppeteer to drive his/her choice of actions (recorded through eg eye-tracking);
4. using machine learning, map these social signals (input state) to the robot behaviours (output state) such that the robot can operate autonomously.

A creative professional (puppeteer, dancer or comedian – corresponding financial compensation is budgeted) will join the group. First she/he will take part to a one-week co-design workshop (4) aiming at finalising the

immersive teleoperation controller and the behaviours of the robot. Then, she/he will interact for about 4 hours a day during a month, with the BRL lab members (200+ researchers). She/he will do so by remotely operating the robot (5) from an (out-of-sight) control room (the BRL CAVE room). The aim will be for the puppeteer to proactively engage with people in the lab, attempting to engage in *social, affective* interactions. This will be achieved by creating/inventing in-situ a new 'grammar' of social behaviour, loosely inspired by those of cats and other pets. These interactions will be fully recorded (including eye-tracking on the puppeeter) (6), in order to create a unique dataset of complex social interactions, suitable for machine learning. The PI has already extensive experience in recording such datasets (see (Lemaignan et al. 2018) for instance).

Over the following four months, a deep neural network will be designed and trained (7) for the regression task of generating continuous social behaviours from perceived social signals. In parallel, a software controller will be developed (8) to enable generic autonomous capabilities (like autonomous navigation) for which the BRL has extensive expertise.

Finally, the last four months will be dedicated to in-situ testing of the autonomous system (9). We will seek to conduct a large scale study within the lab, over a period of several weeks. For this study, the robot is expected to be fully autonomous. However sufficient amount of time is planned for additional iterations on the development of the robot controller if deemed necessary. We aim at publishing the results of this main study shortly after the end of the one-year period.

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 need treatments (eg [36]). 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.

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

WP4 is the technical core of the project: we will create a novel socio-cognitive architecture for the robots, bringing together advanced perception of the human social dynamics, intrinsic motivation to support human interactions, and human-in-the-loop machine learning to create transparent, trustworthy action policies. This WP is high-risk/high-gain, as no such combined approach has been successfully implemented and deployed in real-world, complex social situations. I mitigate the risk by ensuring cognitive functions are decoupled from each other where sensible, and in particular, by ensuring that the robot actions are generated independently through both an intrinsic motivation mechanism, and a human-taught machine learning action policy, hence creating a level of cognitive redundancy (with the corresponding arbitration mechanisms in place where necessary).

Disembodied Cognitive Architectures Reviews: [24], [88] – extended in [42], [30], [44], [82], [84].

Social Robotics Architectures *Society of Mind*-inspired paradigm underpinning practical architectures
Mostly functional architectures: integration models rather than cognitive architectures per se.
ACT-R/E [86], HAMMER [27], PEIS Ecology [69, 26], CRAM/KnowRob [15, 83], KeJia [23]

To be sorted... To check – they may or may not be all relevant to sHRI.
Architectures that...

- ...reason about other agent mental state
  - Scone [31]
  - Polyscheme [16]
- ...'represent knowledge'
  - [97]
- ...cover the 'whole interaction stack'
  - POETICON++ arch [2]

T4.2

The current state-of-the-art is limited in this respect: sub-symbolic approaches focus on identifying low-level social signals (for instance, gazing, gesture recognition) while symbolic approaches have focused on classical symbolic problems like language understanding. Some attempts have been made to bridge both (for instance,

multi-modal dialogue [54, 56]), but no work to date has been able to interpret the social dynamics themselves (for instance, *are the partners getting along well?*, *Is there any frustration?*, *Are they collaborating or not?*, *Are they engaged in their joint task?*). This is likely due to the complexity of the problem, mixing sub-symbolic and symbolic reasoning in intricate ways. Understanding social dynamics in naturalistic conditions entails identifying a complex net of overlapping social cues, at multiple time scales. Reasoning about these ever-changing, sometimes contradictory, dynamics, is an essential skill for an artificial socio-cognitive system, yet an essentially open research question.

I have recently introduced a dataset of social interaction [52] 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 [12]. This dataset and the related on-going scientific investigations will form the starting point of my approach to this challenge.

T4.1 – A social teleology for robots The case for *teleological* (ie goal-driven) robotic architectures has been made in the past [96], but only effectively realised for relatively simple cognitive systems (like curiosity-driven robot animals [61] or motor babbling in infant-like robots [34]). Socially-driven robots, participating in complex interactions with humans, have been barely investigated. This task covers the overall design of the architecture.

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 [73, 74, 93], this task implements the learning mechanics (including the critical aspect of the interface with the human teacher) to allow human participants to progressively teach the robot a social policy to become a good social helper.

In addition, this task 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, guaranteeing a level of behavioural transparency for the end-users.

T4.3 – Integrating a socially-driven architecture for long-term interaction The socio-cognitive architecture of WizUs robots builds from the principles (the 'why's?') identified in T1.2, and relies on a combination of socially-driven intrinsic motivation (a *social teleology*, T4.1), and human-in-the-loop machine learning (T4.2) to progressively learn an social policy enabling long-term autonomy. This task focuses on 'bringing the pieces together' in a principled manner.

We will specifically look at the requirement for *long term* autonomy: Over the last two years, we have observed a significant increase of studies involving social robots, deployed in real-world settings (schools, care centres) over relatively long periods of time (up to 2 or 3 months at a time) [43, 46], with some promising results in well defined situations, with pre-defined tasks (for example, learning tasks [74], or 'butler' in a social care facility [39]). More generic (long-term) social autonomy however requires additional, beyond-state-of-the-art research to (1) add a *social motivation* mechanism able to drive the robot's intentions over time. This is specifically investigated in T4.1 and T4.2 above; (2) a level of cognitive redundancy to ensure reliable perception and behaviour generation (addressed by this task, with a dependency on the cognitive functions developed in WP2 and WP3).

Additionally, a critical aspect of task T4.3 is to 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 [58] that enables the re-assessement of a policy (here, our intrinsic motivation) based on previous experience (here, the human-taught policy).

WP5: Evidence-based research: demonstrable usefulness of social robots in real-world, complex scenarios

Critically, the scientific investigation of these questions (*conceptual framing of robot–supported human–human interaction*, *socially–driven cognitive architecture*, *online social situation assessment and social behaviour generation*) has to be grounded in real-world applications. Indeed, two complex, high impact applications are detailed in the project. The first one takes place in Bristol's Children's Hospital: a social robot will be permanently deployed in one of ward, supporting isolated children who suffer long-term conditions, in close cooperation with the hospital staff. The second one involves the deployment of social robots in special need schools (SEN schools) in Bristol. 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.

These two applications, detailed hereafter, aim at supporting socially-vulnerable populations: this entails specific risks (detailed and addressed in the Section 4), but also offers a unique opportunity to demonstrate in a real-world, impactful case, how social robots can effectively have a positive, ethical impact, and support in-fine

stronger, richer human-human interactions.

WizUs aims at developing and demonstrating the need for a global approach to social HRI with ambitious, high-impact, socially meaningful, applications.

The two last application scenarios are ambitious and inherently risky, as they target vulnerable populations. This is an however informed choice: first, we 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, we 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 100+ hospitalised children with long-term conditions, and 50+ SEN-educated children to directly benefit of the project, 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.

probably useful to give specific examples of interventions

Evaluation methods quantitative/qualitative; Metrics like Inclusion of Other in Self task and a Social-Relational Interview [90]

Application 1: Social robots in SEN schools

The first study investigate whether a socially assistive robot can effectively support the development (learning?), 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 on-going collaborations (through a project on child-robot interaction for children with autism at Bristol's Mendip School).

Specifically, the two main questions we seek to investigate are: What are the (social & spatial) 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 project consists in deploying the R1 social robot in a Bristol-based SEN school (the Mendip School), to investigate how the robot can help shaping a spatial and social school ecology that fosters mental well-being, while effectively supporting teachers and students in their learning.

The project will adopt a strong participatory design approach, inspired by Patient and Public Involvement methodologies (PPI [20]), with 2 one-day focus groups organised with the school teachers; and one evening focus group with the school parents, prior to the study.

Following the workshops, the teacher-oriented codesign of the robot's activities and supervision tools (eg to start/stop/pause/resume activities) will be finalised and implemented by the research team.

The school study itself will take place during Y3, with the robot permanently based at the school. During these visits, the robot will take part in the regular teaching and other daily routines of the school, and will directly interact with the children. While the robot's behaviours will have been co-designed with the teachers, the robot will be teleoperated by a member of the research team, to ensure we do not create an additional burden for the staff.

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 [90]), 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.

Application 2: A robot companion at the children's hospital

T5.2 – creation and deployment of a small robot companion to support isolated children during their hospital stay, fully integrated and aware of the wider hospital ecosystem. Over the course of this second, one-year long (Y4) experiment, we will deploy one WizUs robot at the Bristol Children's Hospital. Using a *mutual shaping* approach [95] to design the role of the robot with the different stakeholders (nurses, doctors, parents, children), we 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 with a yard.

Specific resources this task will take place at the Bristol Children Hospital. Several preparatory meetings already took place with the head of the hospital education service J. Bowyer, who will support the project, giving

me access to two of the long-term conditions wards for the duration of the studies.

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

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* [92])?
- 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 museum, 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 [80, 22], 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 [59, 77]. As an example, a common objection to human-robot interaction is the perceived deceptive nature of the robot's role. It has been argued [19] 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 [76], 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 [50, 52, 74]), 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. 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 Privacy' document.

The project will also follow the European Commission recommendations for Responsible Research and Innovation (RRI). RRI is defined in [81] (and has been subsequently adopted by the UK Engineering and Physical Sciences Research Council [62]) 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 [79] [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 (museum 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 museum 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 spato-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 [55, 70, 51]. As such, T2.1 can be described as **low-risk, medium-gain**.

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

Tasks 2.2, 2.3, 2.4 Work on real-time modeling of social dynamics in real-world environments are only beginning 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 [52, 71], and does not foreseen major difficulties. The resulting components have however the potential to unlock a new class of social robots, aware in real-time of their social surroundings and dynamics. These tasks are thus considered **low-risk, high-gain**.

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

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

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

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

Task 4.2 The techniques developed in T4.2 have been previously used and tested by the PI in two different real-world environments [74, 93]. 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 [56]. 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 [56]. 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 refugee camps) research stay including engagement with the staff/charities and the children themselves; use of art-based participatory techniques like puppetering. I will also perform field testing early on in the project, 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.

The PI, Pr. Séverin Lemaignan, has been working for 12+ years in human-robot interaction, and over the last 6 years, specifically in the field of child-robot interaction. His profile is both highly technical, with hundreds of hardware and software contributions to the worldwide robotic community; and highly experimental, running dozens of studies and experiments with children, including long-term ones, in multiple schools and in healthcare environments. He is in a unique position to deliver both a scientific breakthrough on social situation assessment for robots, and a high-impact societal change.

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] K. Baraka, P. Alves-Oliveira, and T. Ribeiro. "An extended framework for characterizing social robots". In: *arXiv preprint arXiv:1907.09873* (2019).

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

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

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

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

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

[12] 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).

[13] 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.

[14] 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.

[15] 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.

[16] P. Bello. "Shared representations of belief and their effects on action selection: A preliminary computational cognitive model". In: *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society*. 2011, pp. 2997–3002.

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

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

[19] 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.

[20] 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.

[21] 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).

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

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

[31] S. E. Fahlman. "Using Scone's Multiple-Context Mechanism to Emulate Human-Like Reasoning." In: *AAAI Fall Symposium: Advances in Cognitive Systems*. Vol. 11. 2011, p. 01.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

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

[38] S. Harnad. "The symbol grounding problem". In: *Physica D: Nonlinear Phenomena* 42.1 (1990), pp. 335–346.

[39] 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.

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

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

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

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

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

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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).

[51] 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.

[52] 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.

[53] 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.

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

[55] 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.

[56] 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.

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

[58] 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.

[59] 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.

[60] D. P. Oswald and T. H. Ollendick. "Role taking and social competence in autism and mental retardation". In: *Journal of autism and developmental disorders* 19.1 (1989), pp. 119–127.

[61] 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.

[62] 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.

[63] A. Özgür, S. Lemaignan, W. Johal, M. Beltran, M. Briod, L. Pereyre, F. Mondada, and P. Dillenbourg. "Cellulo: Versatile Handheld Robots for Education". In: *Proceedings of the 2017 ACM/IEEE Human-Robot Interaction Conference*. 2017. doi: 10.1145/2909824.3020247.

[64] 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.

[65] 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.

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

[67] A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor. ""Alexa is my new BFF" Social Roles, User Satisfaction, and Personification of the Amazon Echo". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2017, pp. 2853–2859.

[68] 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.

[69] 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.

[70] 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.

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

[72] A. Sciuto, A. Saini, J. Forlizzi, and J. I. Hong. ""Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. 2018, pp. 857–868.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

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

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] M. Tonkin, J. Vitale, S. Herse, M.-A. Williams, W. Judge, and X. Wang. "Design Methodology for the UX of HRI: A Field Study of a Commercial Social Robot at an Airport". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '18. Chicago, IL, USA: Association for Computing Machinery, 2018, pp. 407–415. isbn: 9781450349536. doi: 10.1145/3171221.3171270. url: https://doi.org/10.1145/3171221.3171270.

[86] 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.

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] B. Wrede, K. Rohlfing, J. Steil, S. Wrede, P.-Y. Oudeyer, and J. Tani. "Towards robots with teleological action and language understanding". In: *Humanoids 2012 Workshop on Developmental Robotics: Can developmental robotics yield human–like cognitive abilities?* Ed. by Ugur, Emre, Nagai, Yukie, Oztop, Erhan, Asada, and Minoru. Osaka, Japan, Nov. 2012. url: https://hal.inria.fr/hal-00788627.

[97] S. Zhang, M. Sridharan, M. Gelfond, and J. Wyatt. "Towards an architecture for knowledge representation and reasoning in robotics". In: *Social Robotics*. Springer, 2014, pp. 400–410.

B2.c Description of resources

Requested budget

This table does not count toward the page limit and is to be provided online

Research team and PI commitment

Table 6.1 provides an overview of the time allocation per members of the team, over the course of the project.

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 (senior 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 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 museum (to this end, the museum 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 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 (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, 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.

Subcontracting

The subcontracting amount covers the specific content creation and public communication costs, required to integrate the robot in the Bristol Science museum.

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.

Labour cost		**€ 1,413,344**
Severin Lemaignan	36	€ 287,455
RF 1 (sociology/anthropology, WP1)	36	€ 195,779
RF 2 (social signal processing/machine learning, WP2)	48	€ 263,722
RF 3 (cognitive architecture for robotics, experiments, WP4 WP5)	60	€ 333,050
PhD 1 (social signal processing/machine learning, WP4 WP5)	42	€ 64,233
RF 4 (behaviour generation, WP3)	48	€ 269,105
Travel		**€ 65,865**
Conferences (1 per person per year): Registration		€ 22,910
Conferences (1 per person per year): Travel		€ 42,955
Equipment		**€ 334,527**
R1 Robot (inc VAT)		€ 292,800
R1 Shipping		€ 7,000
Robot Installation and Training		€ 3,800
Workstation suitable for machine learning		€ 22,910
Sensors (inc RGB-D cameras, one eye tracker)		€ 8,018
Materials		**€ 11,455**
Other Consumables		€ 11,455
Other		**€ 23,455**
Open access fees (2 Article per year)		€ 12,000
Participants compensations		€ 3,436
Data Storage		€ 2,291
Cloud Computing		€ 5,727
Audit		€ 5,000
Total		**€ 1,853,645**
Indirect		**€ 463,411**
EC Contribution		**€ 2,317,056**

	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 (WP4, WP5)	1	1	1	1	1	60
PhD 1 (WP4, WP5)		1	1	1	0.5	42
Post-doc 4 (WP3)		1	1	1	1	48

Table 6.1: Full-time equivalent for the research team members

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.

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.

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's multidisciplinary approach aims to create autonomous devices capable of working independently, with each other, or with humans. BRL draws on robotics, electrical & mechanical engineering, computer science, psychology, cognitive science and sociology. 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: HRI, collective robotics, aerial robotics, neuro-inspired control, haptics, control systems, energy harvesting and self-sustaining systems, rehabilitation robotics, soft robotics and biomedical systems. 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.

Hardware-wise, the Bristol Robotics Lab, where this fellowship will take place, offer unique support and facilities for robotic hardware development: the laboratory's dedicated equipment includes two industrial-grade rapid prototyping machine, a laser cutter, one 5-axis digital milling machine, all the required facilities for PCB prototyping, and a team of six full-time technicians, specialised in hardware development. The BRL has indeed 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 implementation of the WizUs robots. This will be directly supported by dedicated technician time, costed in the project.

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 a coherent and strong pathway to impact to the project, including further engagement with industrial partners and spin-off opportunities.

Other in-kind contributions

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