Towards "Machine-Learnable" Child-Robot Interactions: the PInSoRo Dataset

Séverin Lemaignan[1], James Kennedy[1], Paul Baxter[2] and Tony Belpaeme[1]

***Abstract*— Child-robot interactions are increasingly being explored in domains which require longer-term application, such as healthcare and education. In order for a robot to behave in an appropriate manner over longer timescales, its behaviours should be coterminous with that of the interacting children. Generating such sustained and engaging social behaviours is an on-going research challenge, and we argue here that the recent progress of deep machine learning opens new perspectives that the HRI community should embrace. As an initial step in that direction, we propose the creation of a large open dataset of child-robot social interactions. We detail our proposed methodology for data acquisition: children interact with a robot *puppeted* by an expert adult during a range of playful face-to-face social tasks. By doing so, we seek to capture a rich set of human-like behaviours occurring in natural social interactions, that are explicitly mapped to the robot's embodiment and affordances.**

I. Machine Learning: the Next Horizon for Social Robots?

While the family of *recurrent neural networks* have repeatedly made the headlines over the last few years with impressive results, notably in image classification, image labelling and automatic translation, they have been largely ignored in many other fields so far as they are perceived to require very large datasets (hundreds of thousands to millions of observations) to actually build up useful capabilities. Even though neural networks have demonstrated compelling results in open-ended, under-defined tasks like image labelling, they did not stand out as attractive approaches to problems involving high dimensions with relatively small datasets available – like human-robot social interactions.

Besides, if one considers "social interactions" to also entail joint behavioural dynamics, and therefore, some sort of temporal modeling, neural networks look even less enticing as time is notably absent from most of the tasks which neural networks have been successful at.

In 2015, the Google DeepMind team demonstrated how a convolutional recurrent neural network could learn to play the game Break-Out (amongst 48 other Atari games) by only *looking* at the gaming console screen [1]. This result represents a major milestone: they show that a relatively small sample size (about 500 games) is sufficient for an artificial agent to not only learn how to play (which requires an implicit model of time to adequately move the Break-Out paddle), but to also create gaming strategies that *look like* they would necessitate planning (the system first breaks bricks on one side to eventually get the ball to break-out and reach the area *above* the remaining bricks, therefore ensuring rapid progress in the game). We argue that the complexity of mechanisms that such a neural network has been able to quickly uncover and model should invite our community to question its applicability to human-robot interactions (HRI) in general, and sustained, natural child-robot interactions in particular.

However, the lack of a widespread HRI dataset suitable for the training of neural networks is a critical obstacle to this initial exploration. Therefore, as a first step, we propose a design for such a dataset, as well as a procedure to acquire it. We hope that discussions during the workshop may help in further refining this proposal.

II. Machine learning and social behaviour

Using interaction datasets to teach robots how to socially behave has been previously explored, and can be considered as an extension of the traditional learning from demonstration (LfD) paradigms to social interactions (for instance [2], [3]). Previous examples have generally focused on low-level recognition or generation of short, self-standing behaviours, including social gestures [4] and gazing behaviours [5].

Based on a human-human interaction dataset, Liu *et al.* [6] have investigated machine learning approaches to learn longer interaction sequences. Using unsupervised learning, they train a robot to act as a shop-keeper, generating both speech and socially acceptable motions. Their approach remains task-specific, and while they report only limited success, they emphasise the "life-likeness" of the generated behaviours.

Kim *et al.* [7] highlight that applying deep learning to visual scene information in an HRI scenario was successful, but that generating behaviours for the robot to be able to act in a dynamic and uncertain environment remains a challenge.

These examples show the burgeoning interest of our community for the automatic learning of social interactions, but also highlight the lack of structure of these research efforts, as further illustrated by the quasi-absence of public and large datasets of human-robot interactions. To our best knowledge, only the H^3R Explanation Corpus [8] and the Vernissage Corpus [9] have been published to date. The H^3R Explanation Corpus is a human-human and human-robot

This work has been partially supported by the EU H2020 Marie Sklodowska-Curie Actions project DoRoThy (grant 657227), the EU FP7 DREAM project (grant 611391), and the EU H2020 L2TOR project (grant 688014).

[1]Séverin Lemaignan, James Kennedy and Tony Belpaeme are with the Centre for Robotics and Neural Systems, Plymouth University, U.K. firstname.surname@plymouth.ac.uk

[2]Paul Baxter is with the Lincoln Centre for Autonomous Systems, School of Computer Science, University of Lincoln, U.K. pbaxter@lincoln.ac.uk

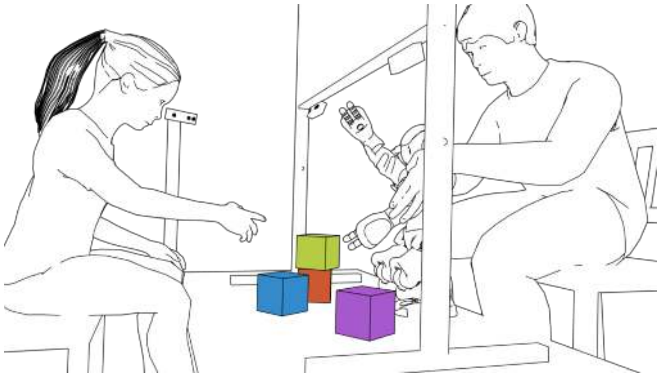

Fig. 1. The acquisition setup: a child interacts with a robot in a range of interactive tasks. The robot is physically guided by an adult expert. We record, in a synchronised manner, the full joint-states of the robots, the RGB and depth video stream from three perspectives (global scene and each of the participant faces), and the sounds (notably, the verbal interactions between the participants).

dataset focusing on a "assembly/disassembly explanation" task and includes physiological signals (22 human-robot interactions), but is not publicly available. the Vernissage Corpus includes one museum guide robot interacting with two people (13 interactions in total), with recordings and annotations of poses and speech audio (stated to be publicly available). Both these corpora are however too small for machine-learning applications.

III. The Plymouth Interacting Social Robots Dataset (PInSoRo)

A. High-Level Aims

The Plymouth Interacting Social Robots (PInSoRo) Dataset is intended to be a novel dataset of human-guided social interactions between children and robots. Once created, we plan to make it freely available to any interested researcher.

This dataset aims to provide a large record of social child-robot interactions that are *natural*: we aim to acquire robot behaviours through corresponding human social behaviour. To this end, we propose that an expert adult will *puppet* a passive robot (Fig. 1). As such, the gestures, expressions and dynamics of the interaction are defined and acted by a human, but as he/she uses the robot body to actually perform the actions, the motions are implicitly constrained by (and thus reflect) the robot embodiment and affordances.

The interactions are supported by a range of short social tasks (described in Section III-B). Critically we propose to limit these tasks to *face-to-face* social interactions, either dyadic or triadic. This constrains the dataset to a more tractable domain, and should ensure technical feasibility. The tasks have to fulfil several key requirements:

- be *fundamentally social*, *i.e.* these tasks would make little or no sense for an agent alone;
- foster rich *multi-modal interaction*: simultaneous speech, gesture, and gaze behaviours are to be observed;
- exhibit *non-trivial dynamics*, such as implicit turn-taking;
- should cover a *broad range of interaction contexts* and situations.

While the tasks will initially be short (in order to acquire a diverse enough dataset), we believe that the captured social behaviours could also be used to inform long-term child-robot interaction. Indeed, naturalistic, rich and socially-oriented multimodal behaviour (beyond simple stereotyped and reactive behaviour) sets the expectation in the human that long-term interactions and social presence [10] can be supported by the robot. Furthermore, we expect such a dataset to allow researchers to uncover several implicit and/or micro-behaviours that, while essential for long-term natural interactions, are difficult to explicitly characterise, and therefore difficult to implement.

B. Tasks

We suggest an initial set of four tasks, lasting about 10 minutes each. They involve collaborative manipulation of simple objects (such as toy cubes), (acted) storytelling, and dialogue-based social gaming. The tasks are intended to be sufficiently different from one another in order to collect a variety of different behaviours, and to minimise task-dependency of the behaviours eventually learnt from the dataset. Physical manipulation of objects across the tasks is limited by the Aldebaran Nao grasping capabilities; the tasks are designed with this in mind, *e.g.* pushing objects away or to the side is possible, whereas pulling them is more difficult.

The tasks are also designed to be playful and engaging, and are derived from classic childrens' games and activities (they are directly inspired by tasks used in other child-robot interaction work, such as [11]). They are thus expected to elicit social interactions that are particularly relevant to child-robot interaction.

a) Task 1: Spatial reasoning: In this task, one partner (child or robot) has a "completed" model made from shapes. Their role is to explain to the other partner how to arrange an identical set of shapes in order to re-create the completed model. The partner with the completed model is not allowed to directly touch the shapes. This task is intended to encourage verbal communication and deictic as well as iconic gestures. It is possible to tune the difficulty of the task through, for example, providing multiple pieces with the same colour, or shape. Similar spatial tasks have been used in other HRI experiments both with adults [12] and children [13].

b) Task 2: Storytelling: The second task revolves around storytelling. To provide a context and collaborative element to the storytelling, "Story Cubes" are incorporated into the task. These cubes are like dice, but with pictures in place of numbers; the pictures serve to guide the story. The two partners are asked to invent a story together, and they take turns in throwing one (large, custom-made) die, arranging the new picture into the story line, and proceeding to tell, and act out, the unfolding story. This task is expected to primarily generate verbal interaction, accompanied by iconic gestures.

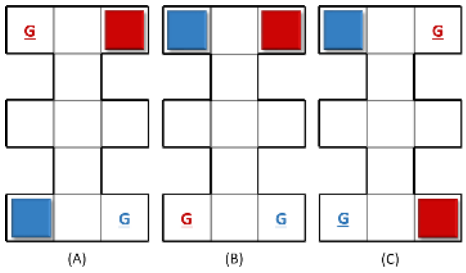

Fig. 2. A sokoban-inspired task requiring collaboration to complete given limitations in robot manual dexterity: the robots face each other across the long edge of each puzzle. Each object (red/blue square) must be pushed to its own goal (red/blue G), in three example levels of difficulty: (A) red and blue objects each simply pushed by one individual, both interactants required, but no explicit collaboration; (B) again a single object requires only a single interactant to manipulate, but some coordination is required due to shared path; (C) each object requires both interactants to manipulate, as well as coordination due to joint path.

c) Task 3: Collaborative strategising: The third proposed task is inspired by the *Sokoban* game (Fig. 2): the two partners must correctly move a set of cubes to locations within a 2D playground by only *pushing* the cubes. Due to the physical setup of the interaction (Fig. 1), the robots are essentially limited to pushing *away* the cubes, transforming the game into a necessarily collaborative activity.

d) Task 4: Party game "Taboo": The fourth proposed task involves triads in a social party game chosen not to require specific gesturing. One such game is "Taboo", a game where one must get others to guess a word without using the word itself. As the game relies only on verbal interaction, we expect all the gestures and gaze behaviour performed by the players to be social backchannel communication, and therefore of direct relevance for the dataset. Using triads is also expected to elicit a richer set of social situations. We expect it to prevent the overfitting of the model to the specific features of dyadic social interactions.

C. Methodology

The envisioned dataset would be comprised of a large number (> 50) of about 30 minutes long recordings of interactions between one child and one puppet-robot, guided by an experimenter (Fig. 1). The pair would be invited to play one or several of the proposed tasks (to be defined after initial pilots). The children would be between 8 and 14 years old. A possibly narrower age range is to be specified once the tasks are precisely defined to ensure the tasks are suitable and engaging for the target age group. Children would typically be recruited from local schools.

We propose to use a Nao robot, and to record the full joint-state of the robot over time. The robot is mostly passive: the feet are firmly fixed on the support table, and all other degrees of freedom, except for the head, are free. The head is externally controlled so that the robot gaze follows the gaze of its human puppeteer in real-time.

The choice of the Nao robot is guided by its small size, making it suitable for puppeting, and its prevalence in the HRI community, resulting in a dataset relevant for a broader academic audience. Also, since Nao is a relatively high degrees-of-freedom (DoF) robot (25 DoFs in total, 5 DoFs per arm), it mimics human kinematics reasonably well. As the motions are recorded in joint space, the dataset can be mapped to other robotic embodiments with similarly configured degrees-of-freedom.

D. Recorded Data

The dataset would comprise the following raw data:

- full 30Hz 25 DoF joint-state of the Nao robot,
- RGB + depth video stream of the scene (see Fig. 1),
- RGB + depth video stream from the child, as seen by the robot,
- speech recording.

Recorded in a fully synchronised manner, these data streams are intended to represent a useful input for many machine-learning techniques. They provide a rich dataset for a range of domains related to social child-robot interaction: from analysis of behavioural alignment between partners (via metrics like the recently proposed *Individual Motor Signature* [14]), to modeling of the dynamics of turn-taking, to the uncovering of implicit in-the-moment synchronisation mechanisms.

This would be complemented by higher-level, post-processed data:

- 68 face landmarks on the child's face, providing options for further facial analysis (like emotion recognition),
- child's skeleton extraction,
- the gaze localisation of each of the participants,
- the 3D localisation of all physical actors (child, all robot parts, cameras, table, manipulated objects),
- the verbal interaction transcripts (automatic transcript with manual verification and correction).

All these sources would be acquired via the ROS middleware (which provides the required mechanism for time synchronisation between the sources) and stored as *ROS bag* files, making it simple to replay the interactions.

As this dataset would contain sensitive data involving children, strict and specific guidelines to ensure the ethical handling of the dataset will be issued before effectively sharing any data.

IV. Discussion

A. Envisioned Applications

The recent advances in machine-learning described in the introduction raise the question of its applicability to the key challenges of artificial intelligence for robotics. Social HRI is a particularly difficult field as it encompasses a large range of cognitive skills in an intricate manner. Application domains of social HRI are typically under-defined, highly dynamic and difficult to predict.

From the data collected, a starting point for machine learning could entail a probabilistic model for reactive behaviours in a given task, *i.e.* finding for each "social cue" the possible set of responses and their probabilities. This could be made generative by using the probability distribution to

seed a roulette-wheel action selection mechanism, effectively creating a probabilistic reactive controller. Whilst simplistic, this is an illustrative example of how the data may be used.

As suggested in the introduction, we also believe that such a dataset could be used to train deep neural networks. While the proposed dataset is very likely not comprehensive enough to train a neural network into an autonomous interactive system, it may be sufficiently rich to train interesting hidden units whose activations would be conditional on specific social situations. For instance, one could imagine that an adequately configured network would generate hidden units able to activate on joint gaze, or on deictic gestures. It must be emphasised that such findings are entirely hypothetical, and we only conjecture them here.

B. Possible Methodological Alternative

Several methodological issues that may impact on the quality of the interaction, the data collection, and the generalisability of results have been anticipated. As the puppeteer behaviours are bound to the embodiment of the robot, it may be that this manipulation inhibits the production of natural behaviours. A small-scale pilot will be used to explore whether or not the puppetted behaviours of the robot inhibit natural interactions with the children.

Besides, one drawback of the proposed acquisition methodology is that the puppeteer remains partially visible to the child (the hands, legs, torso are visible), which may impact the clarity of the interaction (is the child interacting with the robot or with the human behind it?). An alternative acquisition procedure is considered where the puppeteer would remotely control the robot from a different room, using Kinect-based skeleton tracking for the posture control, a head-mounted device for immersive remote vision, and a headset for remote audio. While this adds significant complexity to the acquisition procedure and increases the level of dexterity a task may require, it would provide a cleaner interaction context.

While the tasks have been designed to collect a variety of social behaviours and interaction dynamics, it may be that they are still too similar for any subsequent machine learning to acquire adequately general (*i.e.* not task-specific) behaviours for broader use. Similarly, the use of a single robot may prevent generalisation to other robotic platforms. However, it is not possible to know until algorithms have been applied and tested.

C. Long-Term Considerations

If *useful* social behaviours can be learnt from the initial dataset collected, then this would warrant further collection and exploration of the technique. Transfer to adult-adult pairs could be conducted (possibly with modification of the tasks). Child pairs performing the tasks without the robot could be used to further update behavioural models, as could human behaviours in response to learned robot models, thus providing longer-term adaptivity of behaviour.

Whilst we must acknowledge that the task-centred interactions we propose as part of the PInSoRo dataset are relatively short-term, we do argue that they are capable of simultaneously capturing a range of subtle and complex naturalistic behaviours across a range of different modalities. This type of rich behaviour (by going beyond simple stereotyped and reactive behaviour) supports the expectation in the human that they are interacting with a truly socially competent agent, thus providing the conditions in which long-term child-robot interactions could take place. The application of machine learning algorithms (particularly "deep" methods) provide an opportunity to automatically datamine the solutions to this vastly complex problem that may not be possible with hand-coded systems. Whilst this methodology may yet prove to not be *sufficient* for a complete solution, we propose that the PInSoRo dataset (and others that may follow) establishes a *necessary* foundation for the creation of socially-competent robots over long-term interactions.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[2] C. L. Nehaniv and K. Dautenhahn, *Imitation and social learning in robots, humans and animals: behavioural, social and communicative dimensions*. Cambridge University Press, 2007.

[3] Y. Mohammad and T. Nishida, "Interaction learning through imitation," in *Data Mining for Social Robotics*. Springer, 2015, pp. 255–273.

[4] Y. Nagai, "Learning to comprehend deictic gestures in robots and human infants," in *Proc. of the 14th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2005, pp. 217–222.

[5] S. Calinon and A. Billard, "Teaching a humanoid robot to recognize and reproduce social cues," in *Proc. of the 15th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2006, pp. 346–351.

[6] P. Liu, D. F. Glas, T. Kanda, H. Ishiguro, and N. Hagita, "How to train your robot - teaching service robots to reproduce human social behavior," in *Proc. of the 23rd IEEE Int. Symp. on Robot and Human Interactive Communication*, 2014, pp. 961–968.

[7] K.-M. Kim, C.-J. Nan, J.-W. Ha, Y.-J. Heo, and B.-T. Zhang, "Poro-robot: A deep learning robot that plays video q&a games," in *Proc. of the AAAI 2015 Fall Symposium on AI for Human-Robot Interaction*, 2015.

[8] Y. Mohammad, Y. Xu, K. Matsumura, and T. Nishida, "The H^3R Explanation Corpus human-human and base human-robot interaction dataset," in *Proc. of the Int. Conf. on Intelligent Sensors, Sensor Networks and Information Processing*. IEEE, 2008, pp. 201–206.

[9] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede *et al.*, "The vernissage corpus: A conversational human-robot-interaction dataset," in *Proc. of the 8th ACM/IEEE Int. Conf. on Human-Robot Interaction*. IEEE Press, 2013, pp. 149–150.

[10] I. Leite, C. Martinho, A. Pereira, and A. Paiva, "As time goes by: Long-term evaluation of social presence in robotic companions," in *Proc. of the 18th IEEE Int. Symp. on Robot and Human Interactive Communication*. IEEE, 2009, pp. 669–674.

[11] T. Belpaeme, J. Kennedy, P. Baxter, P. Vogt, E. J. Krahmer, S. Kopp *et al.*, "L2TOR - Second Language Learning Tutoring using Social Robots," in *Proc. of the First Int. Workshop on Educational Robots at the 2015 Int. Conf. on Social Robotics*, 2015.

[12] A. Sauppé and B. Mutlu, "Effective task training strategies for instructional robots," in *Proc. of the 10th Annual Robotics: Science and Systems Conference*, 2014.

[13] C. Zaga, M. Lohse, K. P. Truong, and V. Evers, "The effect of a robot's social character on children's task engagement: Peer versus tutor," in *Proc. of the 2015 Int. Conf. on Social Robotics*. Springer, 2015, pp. 704–713.

[14] P. Słowiński, C. Zhai, F. Alderisio, R. Salesse, M. Gueugnon, L. Marin *et al.*, "Dynamic similarity promotes interpersonal coordination in joint action," *Journal of The Royal Society Interface*, vol. 13, no. 116, 2016.