

Report from Dagstuhl Seminar 2021

Spoken Language Interaction with Virtual Agents and Robots (SLIVAR): Towards Effective and Ethical Interaction

Edited by

Laurence Devillers¹, Tatsuya Kawahara², Roger K. Moore³, and Matthias Scheutz⁴

1 CNRS – Orsay, FR, devil@limsi.fr

2 Kyoto University, JP, kawahara@i.kyoto-u.ac.jp

3 University of Sheffield, GB, r.k.moore@sheffield.ac.uk

4 Tufts University – Medford, US, matthias.scheutz@tufts.edu

Abstract

This report documents the outcomes of Dagstuhl Seminar 2021 “Spoken Language Interaction with Virtual Agents and Robots (SLIVAR): Towards Effective and Ethical Interaction”. Held in January 2020, the seminar brought together world experts on spoken language processing and human-robot interaction. The aims of the seminar were not only to share knowledge and insights across related fields, but also to cultivate a distinct SLIVAR research community. In this report, we present an overview of the seminar program and its outcomes, abstracts from stimulus talks given by prominent researchers, a summary of the ‘Show and Tell’ demonstrations held during the seminar and open problem statements from participants.

Seminar January 5–10, 2020 – <http://www.dagstuhl.de/2021>

2012 ACM Subject Classification Computing methodologies → Natural Language Processing, Computer systems organization → Robotics, Computing methodologies → Philosophical/theoretical foundations of artificial intelligence

Keywords and phrases human-robot interaction, spoken language processing, virtual agents

Digital Object Identifier 10.4230/DagRep.10.1.1

Edited in cooperation with Ali Mehenni, Hugues and Skidmore, Lucy


1 Executive Summary

Laurence Devillers (CNRS – Orsay, FR)

Tatsuya Kawahara (Kyoto University, JP)

Roger K. Moore (University of Sheffield, GB)

Matthias Scheutz (Tufts University – Medford, US)

License  Creative Commons BY 3.0 Unported license

© Laurence Devillers, Tatsuya Kawahara, Roger K. Moore, and Matthias Scheutz

Motivation and aims

Recent times have seen growing interest in spoken language-based interaction between human beings and so-called “intelligent” machines. Presaged by the release of Apple’s Siri in 2011, speech-enabled devices – such as Amazon Echo, Google Home, and Apple HomePod – are now becoming a familiar feature in people’s homes. Coming years are likely to see the



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Spoken Language Interaction with Virtual Agents and Robots (SLIVAR): Towards Effective and Ethical Interaction, *Dagstuhl Reports*, Vol. 10, Issue 01, pp. 1–51

Editors: Laurence Devillers, Tatsuya Kawahara, Roger K. Moore, and Matthias Scheutz



DAGSTUHL
REPORTS Dagstuhl Reports

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

appearance of more embodied social agents (such as robots), but, as yet, there is no clear theoretical basis, nor even practical guidelines, for the optimal integration of spoken language interaction with such entities.

One possible reason for this situation is that the spoken language processing (SLP) and human-robot interaction (HRI) communities are fairly distinct, with only modest overlap. This means that spoken language technologists are often working with arbitrary robots (or limit themselves to conversational agents), and roboticists are typically using off-the-shelf spoken language components without too much regard for their appropriateness. As a consequence, an artefact’s visual, vocal, and behavioural affordances are often not aligned (such as providing non-human robots with inappropriate human-like voices), and usability suffers – the human-machine interface is not “habitable”.

These usability issues can only be resolved by the establishment of a meaningful dialogue between the SLP and HRI communities. Both would benefit from a deeper understanding of each other’s methodologies and research perspectives through an open and flexible discussion. The aim of the seminar was thus to bring together a critical mass of researchers from the SLP and HRI communities in order to (i) provide a timely opportunity to review the critical open research questions, (ii) propose appropriate evaluation protocols for speech-based human-robot interaction, (iii) investigate opportunities to collect and share relevant corpora, and (iv) consider the ethical and societal issues associated with such machines.

Participants

A broad range of expertise was represented by the seminar participants, with a total of 38 attendees including industry experts, PhD students and academics from 14 different countries. The research areas of this interdisciplinary group included SLP, robotics, virtual agents, HRI, dialogue systems, natural language processing, as well as other intersections of SLIVAR.

Seminar overview

The seminar began with short presentations from all attendees, providing them an opportunity to introduce themselves and their research, as well as share their insights on challenges and opportunities in SLIVAR. The presentations were interwoven with four stimulus talks given by leading experts in their respective fields. In light of these presentations, participants formed discussion groups based on the clustering of related topics. The seminar’s schedule was intentionally adaptable to allow for discussions to shift and new groups to form over the course of the week. Alongside discussions, “Show and Tell” sessions were organised to provide participants an opportunity to demonstrate their work and further stimulate discussion.

A non-exhaustive list of topics covered are outlined below along with a selection of the questions discussed within groups.

- Adaptability
 - *How do you cope with the frontier between user adaptation and system adaptation?*
 - *Are there representations that better enable adaptivity to users?*
- Architecture
 - *What are the desiderata for a spoken dialogue system-robot architecture?*
- Ethics
 - *What can we do as scientists and engineers to create ethical agents?*
 - *Should a robot be able to pursue goals that you do not know?*
- Evaluation
 - *How do we evaluate HRI systems effectively and efficiently?*
 - *What are the existing evaluation approaches for SLIVAR?*

- Interaction
 - *How do we bridge the gap between dialogue management and interaction management?*
 - *What kind of interaction modules are useful for dialogue and why?*
- Multimodality
 - *What are the minimum representations units for different modalities?*
 - *What is the added value of multimodal features of spoken interaction in HRI?*
- Natural Language Understanding (NLU) Scalability
 - *How should we approach large scale supervised learning for NLU?*
- Speech in Action
 - *How can we create challenging interaction situations where speech performance is coordinated to a partner's action?*
- Usability
 - *What are the use cases for SLIVAR systems?*
 - *What is the role of physical or virtual embodiment?*

Seminar outcomes

The topics and questions outlined above facilitated a stimulating week of discussion and interdisciplinary collaboration, from which several next steps were established. These include participation in a number of workshops, special sessions and conferences, including but not limited to:

- SIGdial 2020 Special Session on Situated Dialogue with Virtual Agents and Robots ¹
- HRI 2020 Second Workshop on Natural Language Generation for HRI ²
- IJCAI 2020 ROBOTDIAL Workshop on Dialogue Models for HRI ³
- 29th IEEE International Conference on Robot & Human Interactive Communication ⁴
- Interspeech 2020 ⁵

Research and position papers were also discussed, specifically focusing on the evaluation and ethics of SLIVAR systems. For the former, suggestions included a survey of existing evaluation approaches, a report paper on issues in SLIVAR and HRI evaluation, and investigations into the automation of SLIVAR system objective evaluation. For the latter, next steps included a survey of existing architectures for embedded ethical competence and a position paper on ethical machine learning and artificial intelligence.

The final, and perhaps most valuable outcome of the seminar was the establishment of a new SLIVAR community. There was a strong enthusiasm for the discussions during the seminar to continue with a second SLIVAR meeting, as well as suggestions for growing the community through the formal establishment of a special interest group. Overall, the seminar provided a unique opportunity to create a foundation for collaborative research in SLIVAR which will no doubt have a positive impact on future work in this field.

¹ <https://www.sigdial.org/files/workshops/conference21/>

² <https://hbuschme.github.io/nlg-hri-workshop-2020/>

³ <http://sap.ist.i.kyoto-u.ac.jp/ijcai2020/robotdial/>

⁴ <http://ro-man2020.unina.it/>

⁵ <http://www.interspeech2020.org/>

2 Table of Contents

Executive Summary

<i>Laurence Devillers, Tatsuya Kawahara, Roger K. Moore, and Matthias Scheutz</i> . . .	1
---	---

Overview of Stimulus Talks

Ethical Issues in SLIVAR <i>Laurence Devillers</i>	6
Problems and Questions in SLIVAR <i>Tatsuya Kawahara</i>	8
Grounded Language Acquisition for Robotics <i>Cynthia Matuszek</i>	8
Socially Aware Virtual Interaction Partners <i>Catherine Pelachaud</i>	9

Show and Tell

Android ERICA	11
Creating a Voice for the MiRo Biomimetic Robot	11
Incremental Spoken Dialogue and the Platform for Situated Intelligence	12
Furhat – A Social Robot for Conversational Interaction	13
VoxHead	14

Individual Contributions from Participants

Bridging the Habitability Gap <i>Bruce Balentine</i>	14
Ubiquity of Computing and SLIVAR <i>Timo Baumann</i>	16
SLIVAR Needs Models of Interactional Intelligence <i>Hendrik Buschmeier</i>	17
The Role of Social/Moral Norms in SLIVAR <i>Nigel Crook</i>	18
Human-robot Interactions and Affecting Computing: The Ethical Implications <i>Laurence Devillers</i>	19
SLIVAR in Education <i>Johanna Dobbriner</i>	24
Face-to-face Conversation with Socially Intelligent Robots <i>Mary Ellen Foster</i>	25
Building Casual Conversation <i>Emer Gilmartin</i>	26
Architectures for Multimodal Human-Robot Interaction <i>Manuel Giuliani</i>	27
SLIVAR based on Transparency, Situatedness and Personalisation <i>Martin Heckmann</i>	28

On Boundary-Crossing Robots <i>Kristiina Jokinen</i>	31
Human-level Spoken Dialogue Processing for Multimodal Human-Robot Interaction <i>Tatsuya Kawahara</i>	33
Dialogue and Embodiment as Requirements for Understanding <i>Casey Kennington</i>	33
Challenges in Processing Disaster Response Team Communication <i>Ivana Kruijff-Korbayová</i>	35
Personal Statement on SLIVAR <i>Joseph J. Mariani</i>	37
The Importance of Aligning Visual, Vocal, Behavioural and Cognitive Affordances <i>Roger K. Moore</i>	38
Chat, Personal Information Acquisition, and Turn-Taking in Multi-Party, Multimodal Dialogues <i>Mikio Nakano</i>	39
Natural Dialogue Interaction with Autonomous Robots <i>Matthias Scheutz</i>	41
Some Open Questions <i>David Schlangen</i>	41
Interaction Model for SLIVAR <i>Abhishek Shrivastava</i>	43
Personal Statement on Spoken Language Interaction with Virtual Agents and Robots <i>Gabriel Skantze</i>	44
SLIVAR and Language Learning <i>Lucy Skidmore</i>	45
SLIVAR and the Role of the Body <i>Serge Thill</i>	47
What should an agent’s identity be? <i>David R. Traum</i>	48
Are we building thinking machines or are we illusionists? <i>Preben Vik</i>	49
Participants	51

3 Overview of Stimulus Talks

3.1 Ethical Issues in SLIVAR

Laurence Devillers, (CNRS – Orsay, FR)

License  Creative Commons BY 3.0 Unported license
 Laurence Devillers

The new uses of affective social robots, conversational virtual agents, and the so-called “intelligent” systems, in fields as diverse as health, education or transport reflect a phase of significant change in human-machine relations which should receive great attention.

What ethical issues arise from the development of spoken language Interaction with Virtual Agents and Robots (SLIVAR)? Human-chatbot/robot interaction raises the crucial issue of trust, especially for conversational agents who assist vulnerable people. Are nudging machines (SDS) using affective computing and cognitive biases ethical?

The Dilemma of the researchers is on the one hand, to achieve the highest performance with conversational virtual agents and robots (close to or even exceeds human capabilities) but on the other hand, to demystify these systems by showing that they are “only machines”: on the one hand, the designers of conversational agents seek for many to imitate, simulate the dialogical behaviour of humans, on the other hand, users spontaneously anthropomorphise the conversational agents’ capacities and lend them human understanding. The Media Equation [1] explains that people tend to respond to media/computer/robot as they would either to another person by being polite, cooperative, attributing personality characteristics such as aggressiveness, humor, expertise, and even gender depending on the cues they receive from the media. So, an object “which seems to be in the pain”, as the robot Atlas of Boston Dynamics, can inspire some empathy. Asking users not to project human traits on machines is not enough, as some reactions may even appear in spite of this knowledge.

At LIMSI-CNRS, we build agents that can recognise, interpret, process and simulate human language and affect (even a kind of machine-humor). With the capacity of interpretation of the emotional state of humans, a robot can adapt his behaviour and give an appropriate response to these emotions. Naturally, it interacts differently with different individuals. The planned scientific work in our chair HUMAINE focuses on the detection of social emotions in human voice, and on the study of audio and spoken language manipulations (nudges), intended to induce changes in the behaviour of the human interlocutor. A nudge is an indirect suggestion or subtle reminder intended to influence people’s behaviour (Richard Thaler: Nobel Prize in Behavioural Economy, Nov 2017). A “nudge” is a tactic of subtly modifying behaviour of a consumer. Nudging mainly operates through the affective system. Nudges work by making use of our cognitive biases and « irrational » way in decision-making our cognitive capacities are limited, we are lacking self-control, we act emotionally, we act by conformity, we act by laziness, etc.

Nudging could be used in a near future in chatbots and social robots: to incentivise purchase, to influence behaviour that may be and may not be desired by users. The first results from an original pre-experiment, conducted by the proposed HUMAINE Chair’s team in June 2019 in partnership with an elementary school, shows that an AI machine (Pepper robot or Google Home) is more efficient at nudging than adults. Our aim is to study these interactions and relationships, in order to audit and measure the potential influence of affective systems on humans, and finally to go towards a conception of “ethical systems by design” and to propose evaluation measures.

The question of liability arises for designers and trainers of virtual conversational agents. There are several design factors that give rise to ethical problems:

1. Specification problem: a complete specification of a virtual agent is impossible. Laws and rules of conduct in society are formulated in natural language. It is illusory to believe that a complete and precise formulation of natural language elements is possible in a computer language (common sense, irony, culture...).
2. Learning bias: Some of the virtual agent models are trained from data selected by a “coach” (human agent in charge of selecting them). The agent can be discriminating, not fair if the data are badly chosen.
3. Learning without understanding: A virtual agent learns from data, but unlike a human being, he does not understand the meaning of the sentences he generates or perceives.
4. Learning instability: Mistakes are inevitable when a learning system classifies data that do not resemble, or falsely resemble, the data contained in the corpus used for learning. The problem of system robustness is important.
5. Impossible to rigorously evaluate a virtual agent: Dialogue is inherently dynamic. It is difficult to reproduce behaviour or results.
6. Confusion of status: The attribution of a name and a personality to the virtual agent. Maintaining such confusion raises ethical issues. The risk is one of decision manipulation (nudging), isolation, and machine addiction.
7. Trust in virtual conversational agents: Human-agent interaction raises the crucial issue of trust, especially for conversational agents who help vulnerable people. They are currently neither transparent nor evaluated.

It is important to consider the level of trust in a virtual agent, its capabilities and limits and the capabilities and limits of the pair it forms with the user. Some ethical principles have been proposed by the EU experts:

- Beneficence: promoting well-being, preserving dignity, and sustaining the planet
- Non-maleficence: privacy, security and “capability caution”
- Autonomy: the power to decide
- Justice: promoting prosperity and preserving solidarity
- Transparency and Explicability: enabling the other principles through intelligibility and accountability

For example, the objectives of the IEEE SA – P7008 WG which is a working group with public and private partners are:

- understanding human behaviour , nudging and manipulation of choice with spoken dialogue system
- understanding AI nudging applications with public and private partners,
- discussing ethical solutions that guide people to do what’s in their best interest and well-being,
- proposing norms and standards for these ethical solutions.


Conversational virtual agents and robots using autonomous learning systems and affective computing will change the game around ethics. We need to build long-term experimentation to survey Human-Machine Co-evolution and to build ethics by design chatbots and robots.

References

- 1 Byron Reeves and Clifford Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, Cambridge, U.K., 1996.

3.2 Problems and Questions in SLIVAR

Tatsuya Kawahara (Kyoto University, JP)

License  Creative Commons BY 3.0 Unported license
© Tatsuya Kawahara

While smartphone assistants and smart speakers are prevailing and there are high expectations, spoken language interaction with virtual agents and robots (SLIVAR) is not effectively deployed. It is necessary to analyse the reasons and explore use cases. In this talk, first, the dialogue tasks are categorised based on the service types and goals. Next, a variety of social robots and virtual agents are compared according to the affordance. Then, component technologies including ASR, TTS, SLU, dialogue management, and non-verbal processing are reviewed with the focus on critical points for SLIVAR. Finally, evaluation and ethical issues are addressed. Research Questions:

1. Why social robots/agents are not prevailing in society?
2. What kind of tasks are robots/agents expected to conduct?
3. What kind of robots/agents are suitable (for the task)?
4. Why spoken dialogue (speech input) is not working with robots?
5. What kind of other modalities and interactions are useful?
6. What kind of evaluations should be conducted?

3.3 Grounded Language Acquisition for Robotics

Cynthia Matuszek (University of Maryland, Baltimore County, US)

License  Creative Commons BY 3.0 Unported license
© Cynthia Matuszek

For this stimulus talk, Cynthia summarised her current research on grounded language acquisition for human-robot interaction, which she defines as “extracting semantically meaningful representations of human language by mapping those representations to the noisy, unpredictable physical world in which robots operate” [1]. In absence of an abstract, see [2] for her research related to the presentation and [3] for her overview of grounded language acquisition, including future directions and challenges that remain.

References

- 1 Cynthia Matuszek. UMBC, Department of Computer Science and Electrical Engineering; Cynthia Matuszek Bio. <https://www.csee.umbc.edu/~cmat/index.html> Accessed: 03-04-2020.
- 2 Nisha Pillai, Cynthia Matuszek and Francis Ferraro. Deep Learning for Category-Free Grounded Language Acquisition. In *Proc. of the NAACL Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, NAACL-SpLU-RoboNLP, Minneapolis, MI, USA, June 2019. <http://iral.cs.umbc.edu/Pubs/PillaiNAACLws2019.pdf>
- 3 Cynthia Matuszek. Grounded Language Learning: Where Robotics and NLP Meet (early career spotlight). In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI, Stockholm, Sweden, July 2018. http://iral.cs.umbc.edu/Pubs/MatuszekIJCAI2018_earlycareer.pdf

3.4 Socially Aware Virtual Interaction Partners

Catherine Pelachaud (Sorbonne University – Paris, FR)

License © Creative Commons BY 3.0 Unported license
© Catherine Pelachaud

During an interaction, we adapt our behaviours on several levels: we align ourselves linguistically (vocabulary, syntax, level of formality), but also our behaviours (we respond to the smile of our interlocutor, we imitate the posture, the gestural expressiveness...), our conversational strategies (to be perceived more warm or competent), etc. This multi-level adaptation can have several functions: reinforcing engagement in interaction, emphasising our relationship with others, showing empathy, managing the impression we give to others.... The choice of verbal and non-verbal behaviours and their temporal realisation are markers of adaptation. Adaptation, which can take the form of mimicry, synchronisation, alignment, is an important factor in interaction. Several researchers have worked on Embodied Conversational Agents ECAs that can be adapted during an interaction that focuses on imitation, relationship building, empathy [1, 2, 3].

We have conducted several studies to provide the ACA with the capacity for interaction. First, we developed models for agents who were either speakers or interlocutors [4]. Today, we are turning our attention to interaction itself; that is, we are interested in developing agents capable of aligning their behaviours with those of their interlocutors, imitating them, synchronising with them [5]. We are also working to give agents the ability to reason about the expressions they display, to measure their potential impact on their interlocutors. Our current models go beyond our first work on modelling the backchannels of interlocutors [6]. In this first approach, a set of rules specified when a backchannel could be triggered. Then we focused our attention on the ability to equip the virtual agent with the ability to enter into behavioural resonance with his interlocutors. We defined a dynamic coupling model to modulate the level of synchronisation between the agent and his interlocutors [7]. The agent is considered as a dynamic system in constant evolution in real time. The states representing the agent's behaviour can be modified and adapted to allow the emergence of synchronisation between the interlocutors. We have conducted several studies to measure the impact of this motor resonance capacity on the quality of interaction perceived by users. We first evaluated this model between two virtual interactions [8]. We also applied it to the agent capable of laughing and the user listening to funny music (generated according to Peter Schickele's PDQ Bach model) [9]. Virtual agents that are able to synchronise dynamically with other agents or human users are perceived as being more socially involved in the interaction than agents that only send backchannels but do not show any motor resonance.

Now, we focus on the ability to equip the virtual agent with the ability to adapt his behaviour with his interlocutors. We have developed several models that address different aspects of adaptation during an interaction. Over the past two years, we have developed an architecture that allows an ACA to adapt to the non-verbal behaviours of the user during an interaction. We conducted three studies on different levels of adaptation: conversational strategy, nonverbal behaviours, multimodal signals. Each adaptation mechanism has been implemented in the same architecture that includes multimodal analysis of user behaviour using the Eyesweb platform [10], a dialogue manager (Flipper [11]), our virtual agent GRETA. The architecture was adapted to each study. The same scenario was used for the three studies carried out at the Musée des sciences de la Cité des sciences et de l'industrie de Paris. The agent was used as a guide for an exhibition on video games at the Science Museum.

Each of these three studies involved between 70 and 100 participants and followed a similar protocol. Participants first completed a questionnaire based on the NARS questionnaire, used in robotics, to measure their apriori about virtual agents, then interacted with the agent and finally answered other questionnaires on their perception of the agent and interaction. Several hypotheses have been validated, in particular with regard to the competent condition (study 1) and the condition in which the agent adapted his smile to the user's smile (study 3). Study 2 also highlighted the primacy of the warm dimension. In each of the studies, the agent who adapted his or her behaviours to maximise the participants' impression or level of engagement was better perceived. The different coping mechanisms, whether in conversational strategies, non-verbal behaviours or signals, have helped to improve the user experience of the interaction.

References

- 1 Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. Empathy in virtual agents and robots: A survey. In *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):11, 2017.
- 2 Lixing Huang, Louis-Philippe Morency, and Jonathan Gratch. Virtual rapport 2.0. In *International Workshop on Intelligent Virtual Agents*, pages 68-79, 2001. Springer.
- 3 Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*, pages 218-233, 2016. Springer.
- 4 Marc Schroder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, et al. Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165-183, 2012.
- 5 Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69-87, 2012.
- 6 Elisabetta Bevacqua, Maurizio Mancini, and Catherine Pelachaud. A listening agent exhibiting variable behaviour. In Helmut Prendinger, James C. Lester, and Mitsuru Ishizuka, editors, *Proceedings of 8th International Conference on Intelligent Virtual Agents, IVA 2008*, Lecture Notes in Computer Science, volume 5208, pages 262-269, Tokyo, Japan, 2008. Springer.
- 7 Magalie Ochs, Catherine Pelachaud, and Gary Mckeown. A user perception-based approach to create smiling embodied conversational agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):4, 2017.
- 8 Ken Prepin, Magalie Ochs, and Catherine Pelachaud. Beyond backchannels: co-construction of dyadic stance by reciprocal reinforcement of smiles between virtual agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013.
- 9 Maurizio Mancini, Beatrice Biancardi, Florian Pecune, Giovanna Varni, Yu Ding, Catherine Pelachaud, Gualtiero Volpe, and Antonio Camurri. Implementing and evaluating a laughing virtual character. *ACM Transactions on Internet Technology*, 17(1):1-22, 2017.
- 10 Gualtiero Volpe, Paolo Alborno, Antonio Camurri, Paolo Coletta, Simone Ghisio, Maurizio Mancini, Radoslaw Niewiadomski, and Stefano Piana. Designing multimodal interactive systems using eyesweb xmi. In *SERVE@ AVI*, pages 49-56, 2016.
- 11 Jelte van Waterschoot, Merijn Bruijnes, Jan Flokstra, Dennis Reidsma, Daniel Davison, Mariët Theune, and Dirk Heylen. Flipper 2.0. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 43-50, November 2018.

4 Show and Tell

A number of participants were able to share their work through two “Show and Tell” sessions during the seminar. Participants provided videos of robots and virtual assistants in the lab, previews of prototypes and works-in-progress, as well as live demonstrations. Below are some examples of the work that was presented.

4.1 Android ERICA



■ **Figure 1** The android “ERICA”.

Presenter: Tatsuya Kawahara (Kyoto University, JP)

Demonstration: ERICA is an android that can be engaged in human-level conversation including attentive listening and job interview.

Further Information: <http://www.sap.ist.i.kyoto-u.ac.jp/erato/index-e.html>

4.2 Creating a Voice for the MiRo Biomimetic Robot

Presenter: Roger K. Moore (University of Sheffield, GB)

Demonstration: MiRo is the first commercial biomimetic robot to be based on a hardware and software architecture that is modelled on the biological brain. In particular, MiRo’s vocalisation system was designed, not using pre-recorded animal sounds, but based on the implementation of a real-time parametric general-purpose mammalian vocal synthesiser tailored to the specific physical characteristics of the robot. The novel outcome has been the creation of an “appropriate” voice for MiRo that is perfectly aligned to the physical and behavioural affordances of the robot, thereby avoiding the “uncanny valley” effect and contributing strongly to the effectiveness of MiRo as an interactive agent.

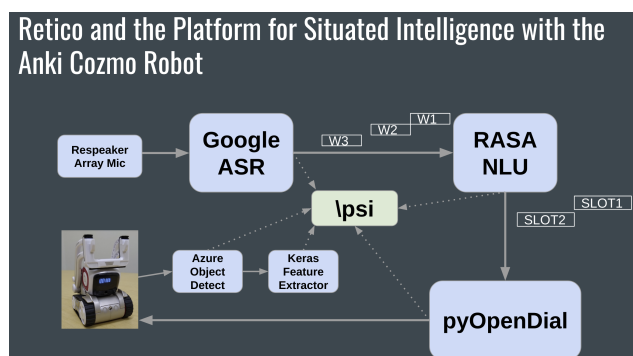


■ **Figure 2** Seminar participants interacting with MiRo.

References

- 1 Roger K. Moore and Ben Mitchinson. Creating a Voice for MiRo, the World's First Commercial Biomimetic Robot. In *Proceedings of INTERSPEECH 2017*, pages 3419–3420, Stockholm, 2017.
- 2 Roger K. Moore and Ben Mitchinson. A biomimetic vocalisation system for MiRo. In M. Mangan, M. Cutkosky, A. Mura, P. F. M. J. Verschure, T. Prescott, and N. Lepora (Eds.), *Living Machines 2017, LNAI 10384*, pages 363–374, Stanford, CA, 2017. Springer International Publishing.

4.3 Incremental Spoken Dialogue and the Platform for Situated Intelligence



■ **Figure 3** Module architecture for voice-controlled navigation.

Presenter: Casey Kennington (Boise State University, US)

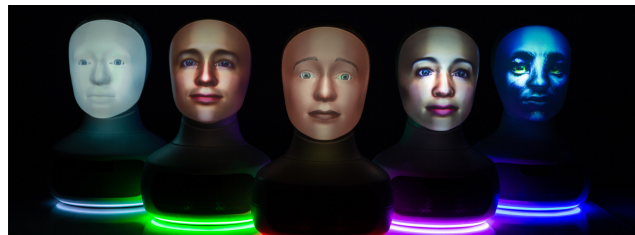
Demonstration: Voice controlled navigation for the Anki Cozmo robot using multimodal, incremental spoken dialogue processing with the Platform for Situated Intelligence and ReTiCo frameworks. Cozmo was able to perform simple navigation commands.

Further Information: <https://anki.com/en-us/cozmo.html>

<https://github.com/microsoft/psi>

<https://github.com/Uhlo/retico>

4.4 Furhat – A Social Robot for Conversational Interaction



■ **Figure 4** The various personas of Furhat.

Presenter: Gabriel Skantze (KTH Royal Institute of Technology – Stockholm, SE)

Demonstration: Furhat started as a research project at KTH and span off into the company Furhat Robotics in 2014. Furhat is a hardware and software platform that allows researchers and developers to build social robotics applications. Hardware-wise, Furhat has a back-projected face that allows for changing the persona of the robot, as well as expressing subtle facial expressions and accurate lip movement. Software-wise, Furhat provides a platform for building multimodal conversational interactions with Furhat. Furhat is being used by many research groups worldwide and by for real-world applications by companies such as Deutsche Bahn (travel information), Merck (medical screening) and TNG (recruitment interviews).

Further Information: www.furhatrobotics.com

References

- 1 Al Moubayed, S., Skantze, G., and Beskow, J. The Furhat Back-Projected Humanoid Head – Lip reading, Gaze and Multiparty Interaction. In *International Journal of Humanoid Robotics*, 10(1), 2013.
- 2 Skantze, G. and Al Moubayed, S. IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA, 2012.
- 3 Skantze, G. Real-time Coordination in Human-robot Interaction using Face and Voice. *AI Magazine*, 37(4):19-31, 2016.

4.5 VoxHead



■ **Figure 5** VoxHead the humanoid robot.

Presenter: Michael C. Brady (American University of Central Asia, KG)

Demonstration: VoxHead is a humanoid robot, developed to be an “open access” research tool for human-robot [speech] interaction. It is composed of 3D printed parts and off-the-shelf components. The idea is that hobbyists and researchers can build these robots for a fraction of the cost of commercial alternatives. Current work involves developing an operating system for the robot that runs interactive dialogues written in “Fluidscript.” This scripting approach is derived from the W3C standard of VoiceXML, and is similar to writing behaviours for today’s Amazon Alexa, except where video input and motor output are both incorporated to specify multimodal interactions.


Further Information: www.fluidbase.com

5 Individual Contributions from Participants

There are several open problems related to SLIVAR still to explore. This section of the report includes statements from attendees providing their perspective on challenges and opportunities within the field.

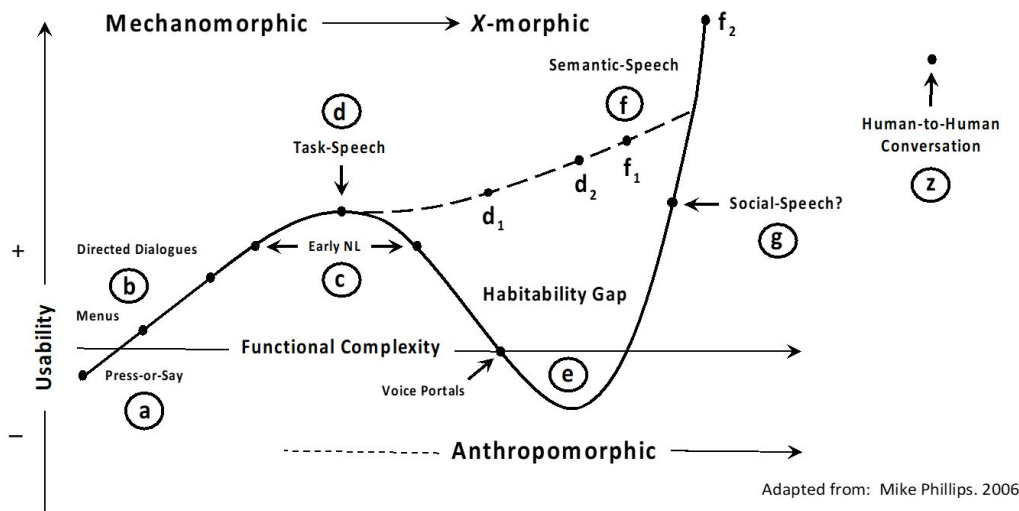
5.1 Bridging the Habitability Gap

Bruce Balentine (Entreprise Integration Group – Zürich, CH)

License  Creative Commons BY 3.0 Unported license
© Bruce Balentine

Basic design philosophy

Anthropomorphism introduces many challenges, among them ethical, uncanny valley, practical implementation and user mind-reading problems. The anthropomorphic goal of “just like human-to-human conversation” (point **z** in Figure 6) leads to the habitability gap via the “death of a thousand cuts.” But what’s the alternative?



■ **Figure 6** Bridging the Habitability Gap.

Mechanomorphism

A mechanomorphic interface presents itself unabashedly as a machine. Such an entity has no mind. It is emotionally agnostic and socially unaware. It is very skilled within its knowledge and task domain, with affordances that convey stability, learnability and discoverability. Mechanomorphism is achieved through the use of:

- Designative meaning in use of language;
- Repetition and reusability;
- Multimodality (tones, visual display, touch and gesture);
- Consistent use of mechanical UI devices (e.g. signposts); and,
- Well-integrated discoverability features.

TaskSpeech (d) is mechanomorphic, and features such affordances. It is built on a smart, half-duplex turn taking foundation. It evolves into (d₁) and (d₂) with full-duplex turn taking, enhanced semantic capabilities and expanded self, user and environment modeling. I am currently involved in a collaboration to develop a TaskSpeech proof-of-concept.

X-Morphism


An X-morphic interface has a mind, allowing (within limits) ostensive-inferential communication and recursive mind-reading skills, including “Theory-of-Mind.” But the system is not conscious, sentient nor sapient, making it an “alien mind” – unlike but compatible with human minds. Since we’ve never built an alien mind before, the X stands for experimental. It is emotionally and socially aware, but not participative. As with mechanomorphism, X-morphic interfaces are task-oriented (their job is to get work done), as measured by traditional HCI usability metrics. The realisation of X-morphism is semantic speech (f) featuring all of the capabilities of TaskSpeech plus extensive negotiation skills, meta-knowledge and meta-cognition, and extended world knowledge within specific domains. (f₂) extends the range of domains. Social speech (g) is semantic speech with full-blown emotional and social competencies.

My roadmap

My trajectory across the habitability gap is: (d) a couple of years, to (f) a decade, to (g) unknown.

5.2 Ubiquity of Computing and SLIVAR

Timo Baumann (Universität Hamburg, DE)

License  Creative Commons BY 3.0 Unported license
© Timo Baumann

Human-computer spoken language interaction has come a long way, and audio-only, task-only virtual agents are carried around by many people on their phones. While some level of functionality for such interactions can now be achieved (as measured by “task success”), interactions with today’s agents are still far from the natural ideal. This is all the more true for spoken language interaction with robots which fight not only the interdisciplinary issues and additional technical complexity involved but also the the unnaturalness of either modality involved and amplified by the combination of these imperfect modalities.

At the same time, computing has become ubiquitous, and cloud-based computing enables us to access our data from an ever growing multitude of devices, from TV and smart speaker via laptop and tablet to smartphones and connected earphones. The advent of 5G networks will help to virtualise even more computation into the cloud while keeping latencies low enough to not be a nuisance in spoken interaction (and robotics). Internet of Things sensors will provide access to all sorts of sensory data. Thus, ubiquitous computing also has the potential to radically improve and change the way that we interact with machines.

These changes come with numerous challenges and opportunities, some of which I try to summarise below:

1. While moving through an ubiquitous computing environment, e.g., leaving the breakfast table, walking down the stairs and to your car, an agent’s realisation should move along and seamlessly transition between devices and modalities, from a possibly embodied agent at the kitchen table to your phone and then your car (or bicycle). The handover poses interesting technical challenges but the more interesting ones are those of availability of modalities (e.g. think twice before reading out e-mails on the subway).
2. Future systems will want to manage and exploit the wealth of data that they acquire about their users from all the modalities and sensors involved. Critical questions here are the users awareness of when she is being observed by the system and how this makes her feel (e.g., does the system observe conversations the user is having with other people?).
3. With the opportunities growing both in system performance and availability, the model of “natural conversation” will limit human-computer interaction. There is no need for a human to be polite to a system, to not interrupt it. Likewise, a system can blend speech, song, signalling noises, etc. in ways that a human never could. It will be interesting to see what forms of “supernatural” sociolect evolve for talking to machines.
4. Human-human interaction, society and culture is easily influenced. Thus, the means of interaction, the assumptions and rules that we design for our future spoken language interaction systems will feed back into human-human language, and from there into society. Already, kids (falsely) ascribe all sorts of properties to Alexa; likewise, female-sounding voices of spoken language systems influence the role model for real(!) females. This influence thus will not only yield an exciting field for research but more importantly requires ethical, societal and cultural far-sightedness from each developer and researcher of spoken language interaction systems.

5.3 SLIVAR Needs Models of Interactional Intelligence

Hendrik Buschmeier (Universität Bielefeld, DE)

License © Creative Commons BY 3.0 Unported license
© Hendrik Buschmeier

Main reference Hendrik Buschmeier: “Attentive Speaking. From Listener Feedback to Interactive Adaptation”. PhD thesis, Faculty of Technology, Bielefeld University, Bielefeld, Germany, 2018.

URL <https://doi.org/10.4119/unibi/2918295>

Language use is successful when we understand the interaction partner and are able to make ourselves understood. Problems in understanding are a major source of frustration in spoken language interaction with virtual agents and robots (SLIVAR), because artificial conversational agents, even in restricted domains, are usually not always able to understand what a human user means – unless users restrict themselves to a specific way of expressing their intention. Although such approaches may work in principle, SLIVAR in this way may feel unnatural and non-spontaneous to users and makes exploration of what a conversational agent can do for the user – discoverability is a general usability-problem in speech-based interfaces – difficult. Both aspects may contribute to the limited acceptance of SLIVAR.

Problems in understanding (non-understanding, partial understanding and misunderstanding) are, however, not limited to SLIVAR. They are prevalent in human communication as well – to an extent that it is argued that “language use is inherently problematic” and that “miscommunication is not a failure but part and parcel of the act of communication” [4, p. 765]. Humans can, however, deal with these problems and repair them interactively through communication.

This insight could be an opportunity for future research on SLIVAR. When problems in understanding the user arise, artificial conversational agents should not give up, but actively try to “come to an understanding” [4, p. 769] with the user by interactively working with them to make themselves better understood. The ability of human interactive language use is, according to Levinson [3], based on the *human interaction engine*, which provides us with *interactional intelligence*.

In this statement, I want to argue, that artificial conversational agents should be endowed with computational models of interactional intelligence, which would allow them to interactively come to an understanding with their interaction partners – in both the speaking and listening roles – and will thus likely be “better” communicators. In previous work, we have computationally modelled a simple form of interactional intelligence based on the dialogue phenomenon of multimodal communicative feedback [1] and could show that an agent equipped with this model communicates more efficiently and that humans rated it more helpful in resolving their understanding difficulties [2].


References

- 1 Hendrik Buschmeier. *Attentive Speaking. From Listener Feedback to Interactive Adaptation*. PhD thesis, Faculty of Technology, Bielefeld University, Bielefeld, Germany, 2018. <https://doi.org/10.4119/unibi/2918295>
- 2 Hendrik Buschmeier and Stefan Kopp. Communicative listener feedback in human-agent interaction: Artificial speakers need to be attentive and adaptive. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pp. 1213–1221, Stockholm, Sweden, 2018.

- 3 Stephen C. Levinson. On the human “Interaction Engine”. In Nick J. Enfield and Stephen C. Levinson, editors, *Roots of Human Sociality: Culture, Cognition and Interaction*, pp. 39–69. Berg, Oxford, UK, 2006.
- 4 Edda Weigand. Misunderstanding: The standard case. *Journal of Pragmatics*, 31:763–785, 1999. [https://doi.org/10.1016/s0378-2166\(98\)00068-x](https://doi.org/10.1016/s0378-2166(98)00068-x)

5.4 The Role of Social/Moral Norms in SLIVAR

Nigel Crook (Oxford Brookes University, GB)

License  Creative Commons BY 3.0 Unported license
© Nigel Crook

We must go beyond simply identifying the ethical issues that arise from spoken language interaction with virtual agents if we are to make those interactions more habitable. In my view, an agent’s (artificial or human) ability to recognise and observe the moral and social norms that surround spoken interaction goes to the heart of what facilitates habitability. This is because these norms embed some core expectations that people have about their interactions with other agents, guiding what it is morally and socially acceptable to say and do and what is not. For example, in human spoken interaction the relative “status” (for want of a better word) of those engaging in dialogue (adult – adult, adult – child, boss – employee, friend – friend, stranger – stranger, salesperson – customer, etc) will strongly influence both what is said and the manner in which it is spoken. The spatial/physical context of the interaction can also set expectations on what verbal interactions are morally and socially acceptable (a conversation in a child’s bedroom, or in a class room, in the office, or at home, etc). More significantly, the cultural context (regional, generational etc) in which the interaction occurs will set the moral and social expectations of the human participants and determine the habitability of that interaction for them. The moment moral or social norms are violated in a spoken interaction, the less habitable that interaction becomes.

Here are some key questions/tasks that I think we need to address here:

1. Identify the role of social/moral norms in spoken language interaction with virtual agents and understand their impact on “habitability”.
This is very challenging as many of these social/moral norms are not directly articulated. They are often acquired by learning through interaction.
2. How can this impact be measured?
I’m not sure there is a metric here – but it may be possible to evaluate how comfortable people are with certain spoken interactions.
3. Explore how virtual agents and robots can be equipped with sufficient social/moral competence to facilitate habitability?
Some work has already been done on this – “top-down” (i.e. rule based) and “bottom-up” (machine learning based) approaches are presented in the literature.
4. Determine the functional aspect of these systems that embody/reveal this social/moral competence to users (e.g. choice of vocabulary, tone of voice, posture, bodily gestures)
Again, difficult to determine how these functional aspects can meet the social/moral expectations of human conversational partners.
5. Accommodating regional/cultural variations in the social/moral norms exhibited through spoken language interaction

It is clear that social/moral norms do not cross cultural boundaries very well. But there are questions here about how these are to be accommodated so that an AI agent can operate in multiple cultural contexts without limiting habitability.

5.5 Human-robot Interactions and Affecting Computing: The Ethical Implications

Laurence Devillers, (CNRS – Orsay, FR)

License  Creative Commons BY 3.0 Unported license
© Laurence Devillers

Social and emotional robotics wants to create companion robots, which are supposed to provide us for example with therapeutic assistance or even monitoring assistance. So, it is necessary to learn how to use these new tools without fear and to understand their usefulness. We need to demystify the artificial intelligence, elaborate ethical rules and put the values of the human being back at the center of the design of these robotic systems. Affective robots and chatbots bring a new dimension to interaction and could become a mean of influencing individuals.

Since the early studies of human behaviour, emotion has attracted the interest of researchers in many disciplines of neuroscience and psychology. Recent advances in neuroscience are highlighting connections between emotion, social functioning, and decision making that have the potential to revolutionise our understanding of the role of affect. Cognitive neuroscience has provided us with new keys to understanding human behaviour, new techniques (such as neuroimaging) and a theoretical framework for their evaluation. The American neuroscientist A. Damasio [1, 2, 3] has suggested that emotions play an essential role in important areas such as learning, memory, motivation, attention, creativity, and decision making. More recently, it is a growing field of research in computer science and machine learning. *Affective Computing* aims at the study and development of systems and devices that use emotion, in particular in human computer and human robot interaction. It is an interdisciplinary field spanning computer science, psychology, and cognitive science. The *affective computing* field of research is related to, arises from, or deliberately influences emotion or other affective phenomena [4]. The three main technologies are emotion detection and interpretation, dialogue reasoning using emotional information and emotion generation and synthesis.

An affective chatbot or robot is an autonomous system that interacts with humans using affective technologies to detect emotions, decide and to simulate affective answers. It can have an autonomous natural language processing system with at least these components: signal analysis and automatic speech recognition, semantic analysis and dialogue policies, response generation and speech synthesis. The agent can be just a voice assistant, a 2D or 3D on-screen synthetic character or a physically embodied robot. Such artefact has several types of AI modules to develop perceptive, decision-making, and reactive capabilities in real environment for a robot or in virtual world for synthetic character. The robot is a complex object, which can simulate cognitive abilities but without human feelings, nor that desire or “appetite for life” that Spinoza talks as *conatus* (effort to persevere in being) which refers to everything from the mind to the body. Attempts to create machines that behave intelligently often conceptualise intelligence as the ability to achieve goals, leaving unanswered a crucial question: whose goals?

The robotics community is actively creating affective companion robots with the goal of cultivating a lifelong relationship between a human being and an artifact. Enabling autistic children to socialise, helping children at school, encouraging patients to take medications and protecting the elderly within a living space is only few samples of how they could interact with humans. Their seemingly boundless potential stems in part from the fact that they can be physically instantiated, i.e., they are embodied in the real world, unlike many other devices. Social robots will share our space, live in our homes, help us in our work and daily life and also share a certain story with us. Why not give them some machine humour? Humour plays a crucial role in social relationships: it dampens stress, builds confidence and creates complicity between people. If you are alone and unhappy, the robot could joke to comfort you; if you are angry, it could help you to put things into perspective, saying that the situation is not so bad. It could also be self-deprecating if it makes mistakes and realises it!

At Limsi-CNRS, we are working to give robots the ability to recognise emotions and be empathetic, so that they can best help their users. We teach them to dialogue and analyze emotions using verbal and non-verbal cues (acoustic cues, laughter, for example) in order to adapt their responses[5, 6]. How are these “empathetic” robots welcomed? To find out, it is important to conduct perceptual studies on human-machine interaction. Limsi-CNRS has conducted numerous laboratory and Ehpad tests with elderly people, or in rehabilitation centers with the Association Approche ⁶, as part of the BPI ROMEO2 project, led by Softbank robotics. Created in 1991, the main mission of the Association Approche is to promote new technologies (robotics, electronics, home automation, information and communication technologies, etc.) for the benefit of people in a situation of disability regardless of age and living environment. We are exploring how the expression of emotion is perceived by listeners and how to represent and automatically detect a subject’s emotional state in speech⁶ but also how to simulate emotion answers with a chatbot or robot. Furthermore, in real-life context, we often have mixtures of emotions [7]. We also conducted studies around scenarios of everyday life and games with Professor Anne-Sophie Rigaud’s team at the Living Lab of Broca Hospital. All these experiments have shown that robots are quite well-accepted by patients when they have time to experiment with them. Post-experimental discussions also raised a number of legitimate concerns about the lack of transparency and explanation of the behaviour of these machines. The winner of the Nobel Prize in economics, the American Richard Thaler, highlighted in 2008 the concept of nudge, a technique that consists in encouraging individuals to change their behaviour without constraining them using their cognitive biases. The behaviour of human beings is shaped by numerous factors, many of which might not be consciously detected. Thaler and Sunstein [8] advocate “*libertarian paternalism*”, which they see as being a form of weak paternalism. From their perspective, “*Libertarian Paternalism is a relatively weak, soft, and non-intrusive type of paternalism because choices are not blocked, fenced off, or significantly burdened*”. Numerous types of systems are already beginning to use nudge policies (ex: Carrot, Canada, for health). Assuming for the time being that nudging humans for their own betterment is acceptable in at least some circumstances, then the next logical step is to examine what form these nudges may take. An important distinction to draw attention to is between “positive” and “negative” nudges (sludges) and whether one or both types could be considered ethically acceptable. The LIMSI team in cooperation with a behavioural economist team in France in the Chair AI HUMAINE HUman-MACHine Affective spoken INteraction & Ethics au

⁶ <http://www.approche-asso.com/>

CNRS (2019-24) will set up experiments with a robot capable of nudges with several types of more or less vulnerable population (children, elderly) to develop nudge assessment tools to show the impact (Project BAD NUDGE BAD ROBOT⁷). The principal focus of this project is to generate discussion about the ethical acceptability of allowing designers to construct companion robots that nudge a user in a particular behavioural direction for different purposes. At the laboratory scale, then in the field, the two teams will study whether fragile people are more sensitive to nudges or not. This research is innovative, it is important to understand the impact of these new tools in the society and to bring this subject on ethics and manipulation by machines internationally⁸. The objects will address us by talking to us. It is necessary to better understand the relationship to these chatty objects without awareness, without emotions and without proper intentions. Users today are not aware of how these systems work, they tend to anthropomorphise them. Designers need to avoid these confusions between life and artifacts to give more transparency and explanation on the capabilities of machines.

Social roboticists are making use of empirical findings from sociologists, psychologists and others to decide their spoken interaction designs, and effectively create conversational robots that elicit strong reactions from users. From a technical perspective, it is clearly feasible that robots could be encoded to shape, at least to some degree, a human companion's behaviour by using verbal and non-verbal cues. But is it ethically appropriate to deliberately design nudging behaviour in a robot?

The imagination of the citizens about robotics and more generally artificial intelligence are mainly founded on science-fiction and myths (*Golem Myth* [9]). To mitigate fantasies that mainly underline gloomy consequences, it is important to demystify the affective computing, robotics and globally-speaking AI science. For example, the expressions used by experts such as “the robots understand emotions”, “they make decisions”, “the robots will have a consciousness”, are not understood as metaphors by those outside the technical research community. The citizens are still not ready to understand the concepts behind these complex AI machines. These emerging interactive and adaptive systems using emotions modify how we will socialise with machines and with humans. These areas inspire critical questions centering on the ethics, the goals and the deployment of innovative products that can change our lives and society [10]. Anthropomorphism is the attribution of human traits, moods, emotions, or intentions to non-human entities. It is considered to be an innate tendency of human psychology. It is clear that the multiple forms of the voice assistants and affective robots already in existence and in the process of being designed will have a profound impact on human life and on human-machine co-adaptation. Human machine co-adaptation is related to how AI is used today to affect people autonomy (in decision, perception, attention, memorisation, ...) by nudging and manipulating them.

Systems have become increasingly capable of mimicking human behaviour through research in affective computing. These systems have provided demonstrated utility, for interactions with vulnerable populations (e.g., the elderly, children with autism). The behaviour of human beings is shaped by several factors, many of which might not be consciously detected. Marketers are aware of this dimension of human psychology as they employ a broad array of tactics to encourage audiences toward a preferred behaviour. One of the main questions in social robotics evaluation is what kind of impact the social robot's appearance has on

⁷ <https://dataia.eu/en/news/bad-nudge-bad-robot-project-nudge-and-ethics-human-machine-verbal-interaction>

⁸ <https://standards.ieee.org/project/7008.html>

the user, and if the robot must have a physical embodiment. The issue is complex, and the Uncanny Valley phenomenon is often cited to show the paradox of increased human likeness and a sudden drop in acceptance. An explanation of this kind of physical or emotional discomfort is based on the perceptual tension that arises from conflicting perceptual cues. When familiar characteristics of the robot are combined with mismatched expectations of its behaviour, the distortion in the category boundary manifests itself as perceptual tension and feelings of creepiness. A solution to avoid the uncanny valley experience might be to match the system's general appearance (robot-like voice, cartoon-like appearance) with its abilities. This can prevent users from expecting behaviour that they will not "see".

Alternatively, users can be exposed to creatures that fall in the uncanny valley (e.g. Geminoids), making the public more used to them. Humans tend to feel greater empathy towards creatures that resemble them, so if the agent can evoke feelings of empathy in the user towards itself, it can enhance the user's natural feeling about the interaction and therefore make communication more effective. Following the reasoning on perceptual categorisation, the robot's appearance as a pleasantly familiar artificial agent and its being perceived as a listening and understanding companion to the user can establish a whole new category for social robots which, in terms of affection and trust, supports natural interaction between the user and the robot.

The research challenge is to build autonomous machines able to learn just by observing the world. For a digital system, autonomy "is the capacity to operate independently from a human operator or from another machine by exhibiting nontrivial behaviour in a complex and changing environment" [11]. In April 2016, Microsoft's Tay chatbot, which had the capacity to learn continuously from its interactions with web users, start racist language after just 24 hours online. Microsoft quickly withdrew Tay. Affective computing and curiosity models will be among the next big research topics. Self-supervised learning systems will extract and use the naturally available relevant context, emotional information and embedded metadata as supervisory signals. Researchers such as A. Bair (MIT lab) created an "Intrinsic Curiosity Model," a self-supervised reinforcement learning system.

The integration of intentionality and human-like creativity is a new area of research. These machines are called "intelligent" because they can also learn. For a robot, the task is extremely difficult because it has neither instinct nor intentions to make decisions. It can only imitate the human being. Giving a robot the ability to learn in interaction with the environment and humans, is the Holy Grail of artificial-intelligence researchers. It is therefore desirable to teach them the common values of life in society. The ability to learn alone constitutes a technological and legal breakthrough, and raises many ethical questions. These robots can be, in a way, creative and autonomous in their decision making, if they are programmed for this. Indeed, according to the American neuroscientist A. Damasio, self-awareness comes from the pleasant or unpleasant feelings generated by the state of homeostasis (mechanisms aimed at the preservation of the individual) of the body. "Consciousness" is a polysemic term: for some, it refers to self-awareness, for others to the consciousness of others, or to phenomenal consciousness, moral consciousness, etc. To be conscious, you need a perception of your body and feelings. The robots would need an artificial body with homeostatic characteristics "similar to ours" to be conscious. The goal of researchers such as K. Man and A. Damasio [12] is to test the conditions that would potentially allow machines to care about what they do or "think". Machines capable of implementing a process resembling homeostasis is possible using soft robotics and multi-sensory abstraction. Homeostatic robots might reap behavioural benefits by acting as if they have feeling. Even if they would never achieve full-blown inner experience in the human sense, their properly motivated behavioural

would result in expanded intelligence and better-behaved autonomy. The initial goal of the introduction of physical vulnerability and self-determined self-regulation is not to create robots with authentic feeling, but rather to improve their functionality across a wide range of environments. As a second goal, introducing this new class of machines would constitute a scientific platform for experimentation on robotic brain–body architectures. This platform would open the possibility of investigating important research questions such as “*To what extent is the appearance of feeling and consciousness dependent on a material substrate?*”

How can we assess a system that learns? What decisions can and cannot be delegated to a machine learning system? What information should be given to users on the capacities of machine learning systems? Who is responsible if the machine malfunctions: the designer, the owner of the data, the owner of the system, its user, or perhaps the system itself? What will be the power of manipulation of the voices of these machines? What responsibility is delegated to the creators of these chatbots/robots?

In my book “Robots and Humans: Myths, Fantasies and Reality” [9], I propose to enrich Asimov’s laws with commands adapted to life-assistant robots. The foundations of these commandments come in part from feedback from experiences of interactions between elderly people and robots. In my new book, “Emotional robots: health, surveillance, sexuality... and the ethics of it all?” (L’Observatoire, March 2020), I describe these “artificial friends” which will take a growing place in society. Just as an airplane does not flap its wings like a bird to fly, we build machines that can imitate without feeling, speak without understanding, and reason without consciousness. While their role can be extremely positive, particularly in the field of health, the risks of manipulation are also real: emotional dependence, isolation, loss of freedom, amplification of stereotypes (80% of these artefacts have voices, names – Alexa, Sofia – and women’s bodies, which turn them into servile assistants or sex robots). Will they be an extension of ourselves? How far will we go to program an emergence of artificial consciousness? Conversational virtual agents and robots using autonomous learning systems and affective computing will change the game around ethics. We need to build long-term experimentation to survey Human-Machine Co-evolution and to build *ethics by design* chatbots and robots.

Developing an interdisciplinary research discipline with computer scientists, doctors and cognitive psychologists to study the effects of co-evolution with these machines in a long-term way is urgent. The machine will learn to adapt to us, but how will we adapt to it? Machines will be increasingly autonomous, talkative and emotionally gifted through sophisticated artificial-intelligence programs?

References

- 1 A. Damasio. *Descartes’ Error: Emotion, Reason, and the Human Brain*. HarperCollins, 1994.
- 2 A. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace, New York, 1999.
- 3 A. Damasio. *Looking for Spinoza: Joy, Sorrow, and the Feeling Brain*. Houghton Mifflin Harcourt, New York, 2003.
- 4 R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, 1997.
- 5 L. Devillers, M. Tahon, M. Sehili, and A. Delaborde. Detection of affective states in spoken interactions: robustness of non-verbal cues, *TAL*, 55(2), 2014.
- 6 L. Devillers, M. Tahon, M. Sehili, and A. Delaborde. Inference of human beings’ emotional states from speech in human-robot interactions. *International Journal of Social Robotics*, 7(4):451–463, 2015.

- 7 L. Devillers, L. Vidrascu and L. Lamel. Challenges in real-life emotion annotation and machine learning based detection, *Neural Networks*, 18(4):407–422, May 2005.
- 8 R. H. Thaler and C. R. Sunstein. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press, 2008.
- 9 L. Devillers. *Des robots et des hommes : Mythes, fantasmes et réalité*. (French) [Robots and Humans: myths, fantasies and reality]. Plon, Paris, France, 2017
- 10 IEEE. Ethically Aligned Design Version 2. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf
- 11 A. Grinbaum, R. Chatila, L. Devillers, J. G. Ganascia. Ethics in Robotics Research: CERN Mission and Context. *IEEE Robotics and Automation Magazine*, 24(3):139–145, 2017.
- 12 K. Man and A. Damasio. Homeostasis and soft robotics in the design of feeling machines, *Nature Machine Intelligence*, 1:446–452, October 2019. <https://doi.org/10.1038/s42256-019-0103-7>

5.6 SLIVAR in Education

Johanna Dobbriner (TU Dublin, IE)

License  Creative Commons BY 3.0 Unported license
© Johanna Dobbriner

The role of robots or virtual agents in their interaction with humans can take many forms, e.g. for production of goods, as a personal assistant, companion, guide, story teller, teacher, and more. Frequently, for that interaction work, spoken language will be involved. Appropriate and engaging dialogue needs to be planned and executed by the virtual agent/robot automatically and in real time.

My own research mainly concerns the use of virtual agents in teaching and education and accordingly both the technical and ethical aspects of that use case are of special interest to me.

Ethical concerns:

- When used to teach children, what ethical constraints are there to consider?
- How can these constraints be built into such a system?
- Are the current measures we take for privacy and data protection sufficient?
- How human-like or realistic should a virtual agent be?

Technical aspects:

- Measuring engagement in real time
- Strategies to keep and increase user engagement
- Customisation of dialogue and interaction
- Generation of realistic conversation
- Integrating multimodal aspects of conversation e.g. speech, intonation, gestures and facial expressions
- Modelling context and situational awareness in the interactions
- Specific needs for dialogue systems in teaching

With researchers from different disciplines involved in SLIVAR coming together, some of the above points may be addressed and discussed from multiple perspectives.

5.7 Face-to-face Conversation with Socially Intelligent Robots

Mary Ellen Foster (University of Glasgow, GB)

License  Creative Commons BY 3.0 Unported license
© Mary Ellen Foster

The overall shape of my research programme is influenced by two well-known facts about interaction:

1. Humans are inherently social creatures who tend to respond socially to any form of technology, whether it is human-like or not [1] This phenomenon is only further enhanced when the technology is embodied in an approximately human-like robot.
2. Face-to-face conversation is the basic form of human communication, and is also arguably the richest possible communications method. Talking to other humans face-to-face permits full, incremental, multimodal communication on all channels – spoken language, prosody, facial displays, proxemics, body posture, body gestures. As Bavelas et al. [2] point out, a useful exercise is to analyse other communication systems by enumerating the ways in which they differ from full face-to-face dialogue.

Taken together, the above facts mean that, when deploying robots into human spaces, it is entirely unavoidable that people will treat those robots as social actors, and will desire and attempt to have a full, multimodal, face-to-face conversation with them. Developers of such robots must acknowledge and understand this phenomenon – and, more importantly, they must also design systems in such a way that the robots are able to detect and respond appropriately to human attempts to engage them in a face-to-face conversation.

More specifically, I am interested in addressing the following research questions in this context:


- What are user expectations of a humanoid robot? How can we understand and manage those expectations (e.g., through appearance, behaviour, or management of the social situation) so that the robots behave in a coherent, predictable way and that users are not disappointed by potentially unrealistic expectations being violated?
- How can a robot allow humans to employ the strategies developed from human-human interaction? More specifically: how can a robot detect and classify the multimodal, conversational social signals of humans in its area? And, how can and should it choose actions to respond appropriately to those signals?
- What are contexts (use cases, scenarios) where a socially intelligent robot can be deployed where it provides an actually useful service for the humans in that space? Many current robot deployment scenarios feel like they were chosen for convenience rather than in response to an actual user demand for a robot to enter those spaces.
- What are the ethical considerations when deploying socially intelligent robots into human spaces? If a robot displays “appropriate” social signals, there is a potential risk of users ascribing intentions and even “feelings” to that robot that are not intended. How can we ensure that these effects are not exploited, especially with vulnerable populations?

References

- 1 Byron Reeves and Clifford Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, Cambridge, U.K., 1996.
- 2 Janet Bavelas, Sarah Hutchinson, Christine Kenwood, and Deborah Hunt Matheson. Using Face-to-face Dialogue as a Standard for Other Communication Systems. *Canadian Journal of Communication*, 22(1), 1997. <https://doi.org/10.22230/cjc.1997v22n1a973>

5.8 Building Casual Conversation

Emer Gilmartin (ADAPT Centre – Dublin, IE)

License  Creative Commons BY 3.0 Unported license
© Emer Gilmartin

While every student of natural language processing is familiar with Jelinek’s 1988 quote “Whenever I fire a linguist our system performance improves”, perhaps not so many know his later⁹ remark on the same topic “It is our task to figure out how to make use of the insights of linguists”. Insights from linguistics, and pragmatics in particular, into social talk should give engineers a basis on which to model such interaction, from knowledge of the kind of data required to features desirable in such talk. Current interest in companion applications and social bots evoke the well-defined natural phenomenon of human social or casual talk. Such “talk for the sake of talking” ranges from short greeting and small talk routines to longer story swapping or discussion, and serves strong social goals – building and maintaining social bonds, entertainment, informing participants of their interlocutors’ personality, values, feelings and affect.

The structure of casual talk involves light chat interleaved with monologic “chunks” of narrative or extended opinion, where one participant dominates, and others contribute feedback. It is clearly not possible to model the entirety of such conversations as a series of adjacency pairs. Chunks are generic, and could be modelled separately, but chat is not easy to specify. Human chat comprises a series of statements and sometimes questions on a topic, interspersed with backchannels and short positive comments with little or no additional information. Contributions often include two elements – a short coda-like acknowledgement of or comment on the previous speaker’s utterance, followed by a new statement or question, either related to the current topic or, when a topic is exhausted, shifting or changing topic. At times of topic exhaustion, conversations have been observed to “idle”, with participants producing only short and generic comments (the first element of the two part contributions described above), for a number of turns until somebody introduces a new topic. These structures reflect the inherent mixed-initiative nature of real conversation, which is thought to involve a levelling of role, power and status differentials between interlocutors, with equally distributed speaker rights.

An example stretch of casual chat is illustrated below:

1. **A:** I saw the match last night.
2. **B:** Oh. Wasn’t Messi great?
3. **A:** Yeah, but Ronaldo really messed up.
4. **B:** Yeah. I guess so.
5. **A:** Yep. He really fluffed that penalty.
6. **B:** Yeah. But a great match overall.
7. **A:** Yeah. It was, wasn’t it?
8. **B:** Yeah, good stuff.
9. **A:** Yeah. Are you golfing much these days?
10. **B:** Yeah, a bit, but not as much as I’d like.

In the snippet above, the two-part contributions can be seen in turns 2, 3, 5, 6, 9 and 10, with a short “backward-facing” segment, followed by a “forward-facing” segment advancing the conversation. Turns 4, 7 and 8 are examples of idling, with low-content segments only. In

⁹ Antonio Zampolli Prize acceptance speech, LREC 2004

the deep learning dialogue modelling community, a number of features have been identified for use in controlling generated talk. Examples are repetition (undesirable), response-relatedness (desirable), and specificity (desirable). The origin of these features is unclear, and may be evidence of a growing “folk pragmatics” in the domain. Considering these three features in the light of pragmatics research, their desirability or otherwise is more nuanced. Repetition is often noted as a failing in S2S models, with systems generating short high frequency low information responses. However, such responses do occur in talk, but they are only the first half of most turns, (or feedback in chunk phases). The challenge is how to generate both elements for realistic social talk. In addition, human interaction involves interlocutor alignment, often posited to be necessary for efficient grounding and processing, and created through lexical and syntactic priming. While, in very short conversations, this may not manifest, once natural conversation gets longer, the level of repetition of words and phrases grows. Thus, a certain level of repetition is necessary and desirable in casual talk. The same factors apply to response-relatedness, as the information rich second element in responses needs to be related to the previous turn while a topic is in progress. However, the introduction of unrelated content for topic changes and even seeming non-sequiters is also common in natural talk. Specificity, or the provision of varied language has been identified as a challenge, with measures of word rarity such as Inverse Document Frequency (IDF) used to boost the level of low frequency vocabulary in generated content. However, language in conversation is generally quite simple, with lower lexical density than in written text. More diverse language is desirable in longer chunks, or in demonstrations of humour or irony. Again, seeming flaws in current methods actually work well where less specificity is needed, in the first part of contributions, but the more content rich second parts need more specificity. A major challenge is measuring success in casual talk, which is not easily definable at the utterance level or even in a single conversation, as the function of the conversation is often a broad longitudinal building of good relationships. Assessment of success, beyond simple measurement of time users spend chatting to the system, is challenging.

5.9 Architectures for Multimodal Human-Robot Interaction

Manuel Giuliani (University of the West of England – Bristol, GB)


License © Creative Commons BY 3.0 Unported license
© Manuel Giuliani

In my work, I am interested in building architectures for multimodal human-robot interaction (HRI). This means architectures that combine verbal and non-verbal input processing components (to understand the human’s utterances) with high-level decision mechanisms (for the robot to decide it’s next actions based on the input) and multimodal output generation (for the robot to respond to the human in a socially appropriate way). There are a lot of challenges related to building these architectures, but specifically I am interested in the following research questions:

- How can we combine sub-symbolic and symbolic components in a hybrid HRI architecture that is modular so that data-driven and rule-based approaches can both be tested?
- Can we build HRI architectures that can be adapted to different usage contexts?
- Are there basic dialogue acts (e.g., for initiating and ending an interaction) that can be reused in many different HRI usage contexts?
- In terms of software engineering, can we build reusable HRI architectures that are based on commonly used middlewares like ROS (Robot Operating System) to allow the community to build on previous work?

5.10 SLIVAR based on Transparency, Situatedness and Personalisation

Martin Heckmann (Honda Research Institute Europe GmbH, DE)

License  Creative Commons BY 3.0 Unported license
© Martin Heckmann

When two humans engage in an interaction, two independent minds with different experiences and views on the world come together. To make this joint activity a success, they have to work together. They need to align their mental representations to be able to form a common ground and define a joint goal for the interaction [1, 2, 3, 4, 5]. The communicative signals they use for such an alignment are not limited to the words they utter but encompasses a multitude of potentially multimodal signals such as prosodic variations and gestures [6, 7]. These signals also help to coordinate when the partners take their turns [8]. The necessary alignment between the partners affects a multitude of domains, the phonological, syntactic and semantic level as well as the situation model [2, 3]. With situation model, I want to refer to different aspects of the current situation such as the environment in which the interaction takes place and the progression of the interaction so far. During the interaction they also make assumptions about their partner's mental world, her ability to perceive, understand and judge, often referred to as common sense, her prior knowledge on the topic of the interaction, her goals and intentions and her current state, traits, skills and personal preferences [4]. The larger the differences between the partners' mental worlds, the more work the alignment process needs. A true alignment can never be reached but is also not necessary as long as the joint goals of the interaction are achieved.

Unfortunately, the different building blocks fundamental to human-human communication which I have outlined above, i.e. the capabilities to perceive and interpret multimodal communicative signals, build an adequate model of the environment, keep track of the history of the interaction, reason with common sense, recruit domain knowledge, interpret the partner's goals as well as model her state, traits, skills and preferences, are at best manifested in a very different form in an artificial agent. In today's artificial agents they are often not present at all or only in a rudimentary form. Furthermore, if and how these capabilities are implemented in the artificial agent is in most cases opaque to the user, i.e. the system is not transparent to the user. This means that the alignment process will take a lot of effort from the human and in the end will leave a large gap between the partners' mental models, often too large to achieve the goal of the interaction.

A common and logical approach to improve the interaction is to improve the system's capabilities, to make them closer to that of a human. Many people have investigated the role of prosody [9, 6] as well as how to integrate the information from all available modalities [10, 11, 12, 13, 14]. We have extended this by multimodal integration and personalisation for prosodic analysis [15, 16] and reference resolution [17]. In classical dialogue systems, the situation model is often limited to keeping track of the dialogue history [18, 19, 20]. For human robot interaction, mainly models have been developed to visually perceive the world and link these percepts to internal concepts, words in particular. This process is often referred to as grounding, sometimes also as anchoring, and can be based on pre-existing categories [21, 22, 23] or learned in interaction [24, 25]. The domain knowledge of classical dialogue systems is typically very narrow, based on hand-crafted domains such as bus information [18], restaurant reservation [19], or, more recently, also technical support dialogues [20]. For the interaction with robots it is common to establish the domain knowledge by combining information from external ontologies with the learning of new representations in interaction

with the user [26, 22]. Common sense reasoning is then implemented by reasoning on these ontologies. We proposed an approach to automatically acquire task-specific domain knowledge from online text-based resources, e.g. in the context of a repair task [27].

Many of the functionalities mentioned above are still nowhere near the level of a human. In my view, the most challenging and promising domains relevant to improve the interaction with artificial agents are better environment models and domain knowledge combined with an adaptation to the individual user.

To really meet human expectations will most likely require an AI with reasoning capabilities on par with a human. Yet such an AI does not seem to be around the corner [28]. Hence, instead of raising the system's perception and reasoning capabilities, I think a more promising approach is to increase its transparency. If the system states, capabilities and intentions are transparent to the human, the alignment process and hence the overall interaction can be much more efficient [29]. Humans use backchannels, facial expressions and emotional displays to give feedback on the interaction and the progress of the alignment [30, 31, 32, 33]. Backchannels, facial expressions and emotions have been also investigated in the context of human-agent interaction [34, 35, 32]. However, the research on emotions in human-agent interaction typically focuses on detecting the emotions of the human to enable emphatic responses of the agent [36, 37]. I consider using emotional displays and other non-verbal cues to provide insights into the capabilities and mental states of the agent an interesting and very relevant topic to help to align the minds of the partners and to improve the interaction. Connected to this is the need to empower the agent to make its reasoning steps transparent, often referred to as explainable AI [38, 39]. We took one step into this direction by suggesting an approach to create annotations with explanations that can be used to train a system to provide explanations for its reasoning [40]. In general, spoken feedback from the agent bears the risk of triggering very high expectations on the human's side with respect to the agent's understanding and reasoning capabilities. I consider finding good models to convey the agent's limitations via its communication an important direction to increase transparency [41].

In short, I am convinced that faster progress can be made if we focus more on making the limitations of the agent transparent than on trying to overcome them.

References

- 1 H. H. Clark and S. E. Brennan "Grounding in communication.," in *Perspect. Soc. Shar. Cogn.*, pp. 127–149, American Psychological Association, 1991.
- 2 M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behav. Brain Sci.*, vol. 27, no. 02, 2004.
- 3 S. Garrod and M. J. Pickering, "Why is conversation so easy?," *Trends Cogn. Sci.*, vol. 8, no. 1, pp. 8–11, 2004.
- 4 K. Friston and C. Frith, "A Duet for one," *Conscious. Cogn.*, vol. 36, pp. 390–405, nov 2015.
- 5 T. Scott-Phillips, *Speaking Our Minds: Why human communication is different, and how language evolved to make it special*. Red Globe Press, 2015.
- 6 J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Commun.*, vol. 36, no. 1-2, pp. 31–43, 2002.
- 7 P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Commun.*, vol. 57, pp. 209–232, 2014.
- 8 H. Sacks, E. A. Schegloff, and G. Jefferson, "A Simplest Systematics for the Organization of Turn-Taking for Conversation," *Language (Baltim.)*, vol. 50, no. 4, p. 696, 1974.
- 9 Y. Sagisaka, N. Campbell, and N. Higuchi, eds., *Computing Prosody*. New York, NY: Springer US, 1997.

- 10 M. Turk, “Multimodal interaction: A review,” *Pattern Recognit. Lett.*, vol. 36, no. 1, pp. 189–195, 2014.
- 11 S. Oviatt, B. Schuller, P. R. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, eds., *The Handbook of Multimodal-Multisensor Interfaces: Foundations, User Modeling, and Common Modality Combinations - Volume 1*. ACM Books, 2018.
- 12 R. Stiefelhagen, C. Fügen, P. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel, “Natural human-robot interaction using speech, head pose and gestures,” in *2004 IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, vol. 3, pp. 2422–2427, 2004.
- 13 K. Funakoshi, M. Nakano, T. Tokunaga, and R. Iida, “A unified probabilistic approach to referring expressions,” in *Proc. 13th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, pp. 237–246, 2012.
- 14 C. Kennington and D. Schlangen, “A simple generative model of incremental reference resolution for situated dialogue,” *Comput. Speech Lang.*, vol. 41, pp. 43–67, 2017.
- 15 M. Heckmann, “Audio-visual word prominence detection from clean and noisy speech,” *Comput. Speech Lang.*, vol. 48, pp. 15–30, 2018.
- 16 A. Schnall and M. Heckmann, “Feature-space SVM adaptation for speaker adapted word prominence detection,” *Comput. Speech Lang.*, vol. 53, pp. 198–216, jun 2019.
- 17 D. Kleingarn, N. Nabizadeh, M. Heckmann, and D. Kolossa, “Speaker-adapted neural-network-based fusion for multimodal reference resolution,” in *Proc. 20th Annu. SIGdial Meet. Discourse Dialogue*, (Stockholm, Sweden), pp. 210–214, Association for Computational Linguistics, 2019.
- 18 J. Williams, A. Raux, D. Ramachandran, and A. Black, “The dialogue state tracking challenge,” in *Proc. SIGDIAL 2013 Conf.*, pp. 404–413, Association for Computational Linguistics, 2013.
- 19 M. Henderson, B. Thomson, and S. Young, “Word-based dialogue state tracking with recurrent neural networks,” in *Proc. 15th Annu. Meet. Spec. Interes. Gr. Discourse Dialogue*, pp. 292–299, Association for Computational Linguistics (ACL), 2014.
- 20 K. Yoshino, C. Hori, J. Perez, L. F. D’Haro, L. Polymenakos, C. Gunasekara, W. S. Lasecki, J. K. Kummerfeld, M. Galley, C. Brockett, J. Gao, B. Dolan, X. Gao, H. Alamari, T. K. Marks, D. Parikh, and D. Batra, “dialogue System Technology Challenge 7,” in *Proc. NIPS2018 2nd Conversational AI Work.*, 2018.
- 21 D. K. Misra, J. Sung, K. Lee, and A. Saxena, “Tell me Dave: Context-sensitive grounding of natural language to manipulation instructions,” *Int. J. Rob. Res.*, 2016.
- 22 S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, “Artificial cognition for social human-robot interaction: An implementation,” *Artif. Intell.*, vol. 247, pp. 45–69, jun 2017.
- 23 L. Fischer, S. Hasler, J. Deigmöller, T. Schnürer, M. Redert, U. Pluntke, K. Nagel, C. Senzel, A. Richter, and J. Eggert, “Where is the tool? – grounded reasoning in everyday environment with a robot,” in *Proc. Int. Cogn. Robot. Work.*, CEUR Workshop Proceedings, 2018.
- 24 D. K. Roy and A. P. Pentland, “Learning words from sights and sounds: a computational model,” *Cogn. Sci.*, vol. 26, pp. 113–146, jan 2002.
- 25 C. Matuszek, “Grounded language learning: Where robotics and NLP meet,” in *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2018-July, pp. 5687–5691, 2018.
- 26 M. Tenorth and M. Beetz, “KnowRob: A knowledge processing infrastructure for cognition-enabled robots,” *Int. J. Rob. Res.*, vol. 32, no. 5, pp. 566–590, 2013.
- 27 N. Nabizadeh, D. Kolossa, and M. Heckmann, “MyFixit : An Annotated Dataset , Annotation Tool , and Baseline Methods for Information Extraction from Repair Manuals,” in *Proc. Twelfth Int. Conf. Lang. Resour. Eval.*, 2020.

- 28 K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, “Viewpoint: When will ai exceed human performance? Evidence from ai experts,” *J. Artif. Intell. Res.*, vol. 62, pp. 729–754, 2018.
- 29 R. K. Moore and M. Nicolao, “Toward a needs-based architecture for “intelligent” communicative agents: Speaking with intention,” *Front. Robot. AI*, vol. 4, p. 66, dec 2017.
- 30 A. Eshghi, C. Howes, E. Gregoromichelaki, J. Hough, and M. Purver, “Feedback in conversation as incremental semantic update,” in *IWCS 2015 – Proc. 11th Int. Conf. Comput. Semant.*, pp. 261–271, 2015.
- 31 K. P. Truong, R. Poppe, I. De Kok, and D. Heylen, “A multimodal analysis of vocal and visual backchannels in spontaneous dialogs,” in *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 2973–2976, 2011.
- 32 C. Lang, S. Wachsmuth, M. Hanheide, and H. Wersing, “Facial Communicative Signals – Valence Recognition in Task-Oriented Human-Robot Interaction,” *Int. J. Soc. Robot.*, 2012.
- 33 P. Ekman, *Emotions revealed: recognizing faces and feelings to improve communication and emotional life*. Times Books, 2003.
- 34 G. Skantze, A. Hjalmarsson, and C. Oertel, “Turn-taking, feedback and joint attention in situated human-robot interaction,” *Speech Commun.*, vol. 65, pp. 50–66, 2014.
- 35 C. Becker, S. Kopp, and I. Wachsmuth, “Why Emotions should be Integrated into Conversational Agents,” in *Conversational Informatics An Eng. Approach* (T. Nishida, ed.), pp. 49–67, John Wiley & Sons, 2007.
- 36 H. Gunes, B. Schuller, M. Pantic, and R. Cowie, “Emotion representation, analysis and synthesis in continuous space: A survey,” in *2011 IEEE Int. Conf. Autom. Face Gesture Recognit. Work. FG 2011*, pp. 827–834, 2011.
- 37 L. Devillers, M. Tahon, M. A. Sehili, and A. Delaborde, “Inference of Human Beings’ Emotional States from Speech in Human–Robot Interactions,” *Int. J. Soc. Robot.*, vol. 7, pp. 451–463, aug 2015.
- 38 A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, “Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda,” in *Conf. Hum. Factors Comput. Syst. – Proc.*, vol. 2018-April, 2018.
- 39 A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
- 40 N. Attari, M. Heckmann, and D. Schlangen, “From Explainability to Explanation: Using a Dialogue Setting to Elicit Annotations with Justifications,” in *Proc. 20th Annu. SIGdial Meet. Discourse Dialogue*, (Stockholm, Sweden), pp. 331–335, Association for Computational Linguistics, 2019.
- 41 R. K. Moore, “Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction,” in *Lect. Notes Electr. Eng.*, vol. 999 LNEE, pp. 281–291, 2017.

5.11 On Boundary-Crossing Robots

Kristiina Jokinen (AIST – Tokyo Waterfront, JP)

License © Creative Commons BY 3.0 Unported license
© Kristiina Jokinen

My work has dealt with communicative competence and enablements of communication that are crucial in modelling natural interaction between humans and between humans and intelligent agents. Within the cascaded dialogue modelling framework, based on the cycle of contact, perception, understanding, and reaction, such enablements are seen as preconditions

for the interaction to proceed in a smooth manner. In particular, my research interests have focused on situational awareness and its signaling by eye-gaze and gesturing. When interacting with social robots, the same enablements are assumed to be valid, albeit with slightly different behaviour patterns, and social robots, enabling spoken interaction with humans, are not only sophisticated computers with a capability to quickly process huge amounts of data, but they are also perceived as interactive agents with an ability to communicate with human partners using natural language. Given the robot's dual characteristics as a computer and as an interactive agent, the main issues in HRI are related to technological enablements to support natural language interaction and to the modelling of complicated issues in the interaction between humans and robots.

Interesting issues in the SLIVAR context concern situational awareness of the agent and the aspects of communication enablements that can be learnt from the human-animal interaction in this respect. Such issues do not only concern speech and vocalisation, not even affect and emotion, but becoming aware of the partner and of the partner being ready for communication. Such recognition of the partner's communicative intention is an essential part of the basic enablements of communication (contact with the partner), and monitoring of the partner is crucial for the dialogue dynamics.


The current HRI and dialogue systems in general lack this kind of situational awareness. They are designed to answer certain factual questions and do some preliminary inferencing, but not to monitor their environment. In my research I have focused on eye-gaze studies and checking the gaze-patterns in various situations from which we can infer the partner's emotional and affective state, as well as comparing gaze patterns between humans and between humans and robots.

From this view point, one of the challenges that I see crucial for the SLIVAR workshop concern timing of the robot's reaction and its coordination with the human partner. How can the dialogue model take into account such immediate reaction and then, deliberate on the higher-level communicative aspects as cooperation, collaboration and planning together with human agents. What kind of dialogue model and interaction architecture would support this kind of functionality and processing of observations.

Another question concerns symbiotic relation between humans and robots. Can robots and robot interaction be as natural as that with pet animals, different but still accepted as one type of the many interactions humans can have with their environment and the world in general. I have introduced the concept of "Boundary Crossing Robot" which refers to robotic agents capable of interacting with humans in everyday life situations, and thus gradually shifting the boundaries of what are typical interactive agents in our environment that we need to take into account and which can take us into account when dealing with everyday tasks. The boundary crossing refers to conceptual boundaries that humans have of the structure of the world, e.g. what are the agents we can communicate with, and thus crossing of the boundary refers to accepting robots as agents which can interact. Another boundary concerns acceptance of robots as partners and co-workers. BCRs are to facilitate interaction and mutual intelligibility between different perspectives among humans and HRI, and especially pave way towards the views of Society 5.0 where the robotic agents and human agents can have symbiotic relation and co-habit the world.

5.12 Human-level Spoken Dialogue Processing for Multimodal Human-Robot Interaction

Tatsuya Kawahara (Kyoto University, JP)

License  Creative Commons BY 3.0 Unported license
© Tatsuya Kawahara

Following the success of spoken dialogue systems (SDS) in smartphone assistants and smart speakers, a number of communicative robots are being developed and commercialised. Since robots have a face and a body, the interaction is essentially multimodal. I have investigated spoken dialogue with robots in the context of multimodal interaction. Compared with the conventional SDSs, people tend to talk to a robot in a closer manner to talking to a human (or a pet?) because of the anthropomorphism and physical presence. This poses fundamental changes in the design and methodology of dialogue and interaction, since the conventional SDSs are designed as a human-machine interface. For example, you don't need a robot just to ask for weather information or news. And a robot should detect when you speak even without pressing a button or saying a magic word.

We first need to explore desirable tasks and interactions conducted by humanoid robots engaged in spoken dialogue. These obviously depend on the character design of the robots, and I focus on long and deep interactions such as counseling and interview, which have a definite task but do not have observable goals. They will expand the potential of communicative robots. Then, we need to enhance the methodology of spoken dialogue processing including speech recognition and synthesis for human-robot interaction. Moreover, non-verbal processing also needs to be incorporated. In particular, smooth turn-taking and real-time feedback including backchannels are critically important for keeping the user engaged in the dialogue, so the interaction will be duplex consisting of not only speaking but also attentive listening.

We are investigating human-level dialogue and developing an autonomous android ERICA. Our ultimate goal is to make the android fully autonomous, passing a “total Turing test”, but we also hope we can make clear what is essential and missing in the current technology for natural dialogue and interaction through this grand challenge.

5.13 Dialogue and Embodiment as Requirements for Understanding

Casey Kennington (Boise State University, US)

License  Creative Commons BY 3.0 Unported license
© Casey Kennington

Symbol grounding for holistic semantics

If machines have any chance of fully understanding humans, they need to be able to learn how to interact with humans *on human terms*; that is, using the most natural and effective communication medium that humans use with each other: speech. Learning speech requires a learning the semantic meaning of language, yet words are not just symbols or high-dimensional vectors, they have connections with the physical (*symbol grounding*) and social (*conversational grounding*) world. A holistic word meaning would, therefore, need to ground into physical modalities such vision, proprioception, interoception, etc., requiring physical embodiment and the ability to interact and manipulate physical objects. Moreover, the *setting* for grounding into the physical modalities is co-located, social interaction with other agents. This physical embodiment in a co-located, interactive speech setting for holistic semantics requires spoken dialogue with robots.

Anthropomorphisation

However, when humans are confronted with robots, they tend to anthropomorphise those robots based on their physical characteristics (e.g., they assign gender and age) and their interactive abilities (e.g., intelligence, or even sincerity). Humans assign affect and valence to robotic behaviours (e.g., a robot with certain face configurations is perceived as *angry*, or a robot that takes a long time to respond is perceived as *confused*). This has implications for the kinds of settings in which they hope to use robots. For example, a robot with a certain color, shape, and facial features will be perceived as friendly or unfriendly in a setting of hospital assistive care. Robots need to have physical characteristics that are amenable to setting in which they are in and what they are tasked with. If the task is to learn semantic concepts (or if learning semantic concepts is a byproduct of another task) the physical characteristics of the robot play a role in the dynamics of the interaction.

Concrete vs. abstract concepts

Concrete linguistics concepts denote physical things; abstract concepts do not, though concrete concepts can be used abstractly (e.g., one can talk about a *dog* without a physical dog being present). Concrete concepts are learned in physical environments, such as referring to physical objects. Abstract concepts are only learned in more social contexts—language building upon language—concepts that are required for holistic semantics, but since concrete concepts are holistically learned in embodied agents, and abstract concepts build on concrete concepts, abstract concepts are likely best learned also in embodied robots.

Social vs. business

Robots and dialogue systems are often task-based with pre-defined goals. Robots and dialogue systems that are optimised for those tasks tend to only have language and affordances (respectively) to complete those tasks. I refer to these as *all business* in that they aren't concerned with socially acceptable (or sometimes required) behaviour. This is of course acceptable because they are machines. Conversely, some systems—robots or dialogue systems—focus more on social aspects, such as social robots or chatbots that don't accomplish any task beyond socialising. I would call these *social* systems. There is a continuum between *social* and *all business* robots and dialogue systems. The point between the two extremes where one decides to set the business/social side depends on the task. Both are required for holistic semantics: building concrete concepts can happen in all business scenarios, but abstract concepts are more likely to come through more socially-driven systems. Moreover, abstract concepts have been shown to be grounded into elements of social dynamics such as valence and affect more than concrete concepts do.

5.14 Challenges in Processing Disaster Response Team Communication

Ivana Kruijff-Korbayová (DFKI – Saarbrücken, DE)

License  Creative Commons BY 3.0 Unported license
© Ivana Kruijff-Korbayová

Abstract

Our current work on team communication processing and teamwork support for disaster response missions provides insights on the communication capabilities robots should ultimately have as team members contributing to such missions. This includes references to complex partially damaged/destroyed dynamic environments with potentially unusual objects, which all pose a challenge for object detection; descriptions of activities and states and knowledge integration over time; it may involve co-present interaction, although more typical is remote interaction, which poses a challenge for orientation and achieving common ground; the teams are complex and teamwork involves multiparty communication, another challenge for common ground modeling; to support shared situation awareness such interaction is often (but not always) multimodal, using a graphical/visual interface, such as a map and/or a video feed.

Emergency first response teams operate in high risk situations and make critical decisions despite partial and uncertain information. In order for technology, such as robots or assistive software agents, to provide optimal support for mission execution, it needs mission knowledge, i.e., run-time awareness and understanding of the current mission goals, team composition, resource allocation, the tasks of the team(s), how and by whom they are being carried out, the state of their execution, etc. Since first responders typically operate under high cognitive load and time pressure, it is paramount to keep the burden of entering mission knowledge into the system at a minimum. It is our goal to develop methods for interpreting the verbal communication in the response team and extracting run-time mission knowledge from it. In [1] we have addressed one particular sub-problem: the recognition of dialogue acts in the communication among the human members of a robot-assisted emergency response team. The acquired mission knowledge is then used to assist the first responders during or after the mission, for example, by supporting the real-time coordination of human and robot actions or by mission documentation generation [2, 3].

The goal we are pursuing is very challenging and requires progress beyond state of the art in many aspects. I describe some of the challenges below.

Noisy speech input

Obviously, one challenge is to deal with noisy speech input. The team members use walkie-talkies and move around in a noisy environment. They may be wearing personal protective equipment. This could have built-in microphones, but currently it does not.

Low-resource domain

There do not exist large amounts of transcribed, let alone otherwise annotated data for disaster response. Recording data in exercises and real missions is only becoming possible with spreading use of digital radio equipment. Obtaining transcriptions and annotations remains a challenge. Obtaining realistic data from robot-assisted missions is even more difficult. Moreover, the content of the team communication varies with the nature of robot involvement in the mission, i.e., the tasks, the degree of robot autonomy and other aspects of the technical system realisation, such as user interface functionality.

Grounding in complex (remote) physical situation and “mixed reality”

The speech refers to the physical environment which is complex, dynamic and nonstandard, e.g., partially destroyed. The team is distributed, which means that they do not fully share the physical/visual context. Moreover, when we deal with robot-assisted disaster response, the human team members are not themselves in the environment, they use robots for remote operation in the danger zone. They share situation awareness in a multimodal way by a combination of speech and a graphical user interface, such as a map with annotations. This means that they refer to objects in the physical world and in the map. We have observed that a kind of mixed or bent reality emerges. Sometimes it is necessary to distinguish the two for proper interpretation, e.g., moving an icon on the map is something else than moving an object in the real world; but sometimes the speech is vague in this respect and the distinction is even not important. BTW, we did not use virtual reality in the experiment we have done so far, but I would expect the worlds to blend even more in that case.

Integrating verbal communication with sensor input

First responders use various sensors for “measuring” the environment, and even more visual and sensor data is available when robots are used. So the verbal communication needs to be together with the visual/sensor input.

Extended/complex “multistep” activities

The activities of the team consist of multiple/many tasks and steps that are related to one another. A point of interest, such as a victim, a hazard source or a fire may be detected by one team and further handled by another team.

Dynamic situation modeling

The physical situation is changing as the mission progresses: resources move around, victims are being extracted, hazard sources neutralised, fires extinguished, etc. The temporal aspect cannot be disregarded in the interpretation of the communication and in the modelling of the situation.

Common ground modelling in multiparty communication

The communication involves multiple team members. Although they normally share the radio communication channel, they explicitly establish pairwise or group threads. Modeling common ground for shared situation awareness needs to take these threads into account and not assume that the team members all have the same mission knowledge.

Overhearing vs. active involvement in communication

Our software agent for interpreting the team communication, as we have conceived it so far, overhears the team communication, it does not itself engage in it as a participant. Nevertheless, there needs to be a possibility to correct or at least reject misunderstandings, i.e., wrong interpretations created by the system, e.g., a wrong or wrongly assigned task. This needs to be done by a non-intrusive way, e.g., by easy to handle text editing in a task management interface, rather than complex clarification dialogues between the system and the user. The functionality we develop for the overhearing agent is a subset of the functionality that will be required when this agent is to actively engage in the team communication.

Communication extended over multiple sessions


Disaster response missions may stretch over extended periods involving multiple sessions with the system or prolonged sessions with team switches. The system needs to handle the integration of knowledge over extended time and support continuing, interrupting and resuming sessions.

References

- 1 T. Anakina and I. A. Kruijff-Korbayov. Dialogue act classification in team communication for robot assisted disaster response. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, Stockholm, Sweden, September 11-13, 2019.
- 2 W. Kasper. Team monitoring and reporting for robot-assisted USAR missions. In *International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, IEEE, pp. 246–251, 2016.
- 3 C. Willms, C. Houy, J.-R. Rehse, P. Fettke and I. A. Kruijff-Korbayov. Team communication processing and process analytics for supporting robot-assisted emergency response. In *International Conference on Safety, Security, and Rescue Robotics (SSRR)*, 2019.

5.15 Personal Statement on SLIVAR

Joseph J. Mariani (CNRS – Orsay, FR)

License  Creative Commons BY 3.0 Unported license
© Joseph J. Mariani

Background

My interest in participating is to have a better view of the state-of-the-art in those fields, and to encourage the two communities of specialists in robotics and human-machine (spoken language) interaction to work more closely, using best practices, and especially evaluation.

State-of-the-art

Where are we now regarding (spoken) interaction with robots? I recently asked Roger Moore at the LT4All conference, who was to say “stone age”?

Mutual understanding

For many years, researchers thought that spoken interaction with robots was simply adding speech input/output to robots. Things have now slightly progressed, but the two communities still have to better understand each other, their research agenda and their constraints.

HRI evaluation metrics and protocols

Objective and quantitative evaluations have been decisive in the development of workable language technologies, starting in the 80s. It seems that the situation is not comparable in the robotics area. How could the evaluation paradigm be also used in robotics? What would be the evaluation metrics and possible protocols? Conducting evaluation tests with embodied robots appears difficult and costly. Can we conduct them in simulated environments, including virtual agents? To what degree is it similar to using actual robots? Similarly, conducting spoken dialogue systems evaluation with human subjects is costly and time consuming. How can we automatise this process with proper data, metrics and protocols?

Multi-party human(s)-robot(s) interaction

Humans are often able to understand whether they are concerned by a statement issued by other humans. How can we handle this when one or several humans are communicating with one or several robots?

Language portability

Most of the developments concern the English language, while robots will be used by humans who want to keep on speaking their native language. What is the size of the effort to port a spoken dialogue system to another language? Do we have to devote the same effort, or is there a way to adapt the system to a different language? It is usually necessary to benefit from large quantities of speech to develop an application, while they are difficult to gather in spoken dialogue situations. How can we address this difficulty?

Ethics of robots

Who is responsible for the robots behaviour, especially given that it is based on machine learning? Is it the designer, the programmer, the trainer, the seller, the owner?

5.16 The Importance of Aligning Visual, Vocal, Behavioural and Cognitive Affordances

Roger K. Moore (University of Sheffield, GB)

License  Creative Commons BY 3.0 Unported license
© Roger K. Moore

Recent years have witnessed astonishing progress in the development of speech-enabled artefacts. Indeed, the appearance of such “intelligent” personal assistants is often hailed as a significant step along the road towards more natural interaction between human beings and future autonomous social agents (such as robots). However, studies into the usage of such technology suggest that, far from engaging in a promised “natural” conversational interaction, users tend to resort to formulaic language and focus on a handful of niche applications which work for them. Given the pace of technological development, it might be expected that the capabilities of such devices will improve steadily. However, evidence suggests that there is a “habitability gap” in which usability drops as flexibility increases. I have hypothesised that the habitability gap is a manifestation of the “uncanny valley” effect whereby a near human-looking artefact (such as a humanoid robot) can trigger feelings of eeriness and repulsion. In particular, I developed a Bayesian model of the effect which reveals that it can be caused by misaligned perceptual cues. So, for example, a device with an inappropriate (e.g. humanlike) voice can create unnecessary confusion in a user. The Bayesian model suggests that the habitability gap can only be avoided if the visual, vocal, behavioural and cognitive affordances of an artefact are aligned. However, given that the state-of-the-art in these areas varies significantly, this means that the capabilities of an artificial agent should be determined by the affordance with the lowest capability, which probably points to an agent’s cognitive abilities as limiting factor. My work in this area suggests that future progress depends on designers taking a whole-system perspective, and that emulating a human being is a recipe for failure. However, I have also raised the possibility that there may be a fundamental limit to the interaction that can take place between mismatched interlocutors (such as humans

and machines), and on-going research is looking into the implications for future speech-based human-robot interaction particularly by studying vocal interactivity in-and-between humans, animals and robots. In summary, I have identified two open challenges in the field of spoken language interaction with virtual agents and robots: (a) how to optimise the multimodal coupling between intentional agents and their environments (including other agents), and (b) how to optimise spoken language understanding between mismatched interlocutors.

References

- 1 Roger K. Moore. A Bayesian explanation of the “Uncanny Valley” effect and related psychological phenomena. *Nature Scientific Reports*, 2(864), 2012.
- 2 Roger K. Moore. From talking and listening robots to intelligent communicative machines. In J. Markowitz (Ed.), *Robots That Talk and Listen* (pp. 317–335). Boston, MA: De Gruyter, 2015.
- 3 Roger K. Moore, Ricard Marxer, and Serge Thill. (2016). Vocal interactivity in-and-between humans, animals and robots. *Frontiers in Robotics and AI*, 3(61).
- 4 Roger K. Moore. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In K. Jokinen and G. Wilcock (Eds.), *Dialogues with Social Robots – Enablements, Analyses, and Evaluation*, (pp. 281–291). Springer Lecture Notes in Electrical Engineering (LNEE), 2016.
- 5 Roger K. Moore and Ben Mitchinson. Creating a Voice for MiRo, the World’s First Commercial Biomimetic Robot. In *INTERSPEECH 2017* (pp. 3419–3420). Stockholm, 2017.
- 6 Roger K. Moore. Appropriate voices for artefacts: some key insights. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, (pp. 7–11). Skovde, Sweden: VIHAR, 2017.
- 7 Sarah Wilson and Roger K. Moore. Robot, alien and cartoon voices: implications for speech-enabled systems. In *1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, (pp. 40–44). Skovde, Sweden: VIHAR, 2017.
- 8 Roger K. Moore and Mauro Nicolao. Towards a Needs-Based Architecture for “Intelligent” Communicative Agents: Speaking with Intention. *Frontiers in Robotics and AI*, 4(66), 2017.
- 9 Roger K. Moore. A “Canny” Approach to Spoken Language Interfaces. In *CHI-19 Workshop on Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions*. Glasgow: ACM, 2019.
- 10 Roger K. Moore. Vocal interactivity in crowds, flocks and swarms: implications for voice user interfaces. In *2nd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2019)*. London, 2019.
- 11 Roger K. Moore. Talking with Robots: Opportunities and Challenges. In *International Conference Language Technologies for All (LT4All)*, Paris, France: Unesco, 2019.

5.17 Chat, Personal Information Acquisition, and Turn-Taking in Multi-Party, Multimodal Dialogues

Mikio Nakano (*Honda Research Institute Japan – Wako, JP*)

License © Creative Commons BY 3.0 Unported license
© Mikio Nakano

Spoken dialogue is one of the promising media for human-machine interfaces. Human users and machines can exchange complicated information through multi-turn dialogues. Recent advances in speech and language processing technologies have enabled us to develop dialogue systems to engage in a variety of tasks. However, only performing task is not enough for

many people to repeatedly use such dialogue systems; it is also crucial to give the user a good impression and establish good a relationship between the user and the system. These are also effective to alleviate the problems caused by the system's intention understanding errors.

Virtual agents and robots (agents hereafter) are expected to play a crucial role in those issues, since they can exploit non-verbal behaviours such as eye gaze, facial expressions, gestures, and postures. However, although there have been many studies on what verbal and physical behaviours of agents can improve the users' impressions, more investigations are needed to establish guidelines for designing agents' behaviours. Below are the challenges that I think are important.

The first is to engage in multimodal chat dialogues, or non-task-oriented dialogues. Chat dialogues have been found effective in giving a good impression [1] and building rapport with users [3, 4], and thus combining task-oriented dialogues and chat dialogues would be effective. However, what kind of non-verbal behaviours should be generated during chat dialogues is yet to be investigated.

Second, acquiring the users' personal information, such as interests, habits, and experiences, through dialogues is one of the important functions for agents, because tailoring dialogues using acquired personal information would contribute to strengthen the relationship between the user and the agent. Not only the linguistic contents of user utterances but also prosody and non-verbal information would be useful for estimating the users' intentions and attitudes during dialogues which leads to precisely acquiring personal information [2].

The third challenge is to achieve smooth turn-taking in multi-party multimodal dialogues. Agents often need to interact with multiple users and sometimes there are multiple agents, but turn-taking in multi-party dialogues is far complicated than that in two-party dialogues [5]. A user speech might be directed to the agent or another user. When a user finishes speaking, the system should start speaking or should wait for another user to start speaking. Non-verbal behaviours of the users and the agents are considered useful for sophisticated conversational floor management.

To address these challenges, recognising and generating social signals in language, prosody, and non-verbal information are important. In addition to investigating what kinds of social signals play crucial roles in these challenges, we also need to make effort for better recognising and generating social signals.

References

- 1 Takahiro Kobori, Mikio Nakano, and Tomoaki Nakamura. Small Talk Improves User Impressions of Interview Dialogue Systems. In *Proc. SIGDIAL-2016*, pp. 370-380, 2016.
- 2 Masahiro Araki, Sayaka Tomimasu, Mikio Nakano, Kazunori Komatani, Shogo Okada, Shinya Fujie, Hiroaki Sugiyama. Collection of Multimodal dialogue Data and Analysis of the Result of Annotation of Users' Interest Level. In *Proc. LREC 2018*, 2018.
- 3 Gale M. Lucas, Jill Boberg, David R. Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski and Mikio Nakano. Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots. In *Proc. HRI-2018*, pp. 344-351, 2018.
- 4 Gale M. Lucas, Jill Boberg, David R. Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski and Mikio Nakano. Culture, Errors, and Rapport-building Dialogue in Social Agents. In *Proc. IVA 2018*, pp. 51-58, 2018.
- 5 Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. Estimating response obligation in multi-party human-robot dialogues. In *Proc. Humanoids 2015*: pp. 166-172, 2016.

5.18 Natural Dialogue Interaction with Autonomous Robots

Matthias Scheutz (Tufts University – Medford, US)

License © Creative Commons BY 3.0 Unported license
 © Matthias Scheutz
URL <https://hrilab.tufts.edu/publications/>

Natural language is often viewed as an appendix or add-on to a robotic control architecture in the robotics community and there is currently still little interest in deep natural language integration. For one, the reason is that robotics tasks are difficult enough as is, and enabling natural language interactions on robots requires not only expertise in various areas of computational linguistics (such as parsing and dialogue systems), but also an understanding of how language interacts with perceptions and actions, and how humans use spoken language in dialogue interactions (which is very different from processing written text). But more importantly, we still do not have good architectural theories of how language needs to be integrated into a cognitive robotic architecture: What roles should prosody and disfluencies play in speech recognition and subsequent parsing? How are referential expressions resolved in open worlds when the robot does not even know the referent? What kind of inferences and common sense knowledge are required for robots to understand indirect requests (expressed as indirect speech acts)? When should the robot hold the floor in dyadic interactions and how should it do that (forget the dynamics of multi-user dialogues with different overlaps among the speakers' utterances)? And how are the natural language processing components in the architecture connected to the rest to allow for information exchanges, the sharing of mutual constraints, the coordination among different parallel processes, and the overall time-sensitive processing required by human speakers? These are just a few of the open questions that need to be tackled if we want to build robots that are reasonably language competent and can be instructed and taught naturally. Critically, we need to revisit component algorithms for all subsystems, from the ASR, to the syntactic and semantic parsers, the pragmatic reasoner, the dialogue manager, the text generation component, and the speech synthesiser and determine the extent to which they can work incrementally in a time-sensitive manner; for that is what humans expect from interlocutors, and this includes non-linguistic aspects such as attention shifts, search and exploration actions (e.g., turn the head, purposefully looking in a particular direction, etc.), carefully timed backchannel feedback while the interlocutor is still speaking, as well as the initiation of actions, including speech actions, while an interlocutor's utterance is still processed.

5.19 Some Open Questions

David Schlangen (Universität Potsdam, DE)

License © Creative Commons BY 3.0 Unported license
 © David Schlangen

What can deep (end-to-end) learning offer to dialogue research?

Deep learning success in many areas (MT, ASR) has so far not translated into success in dialogue modelling. There are some advances in NLU (intent classification), but also, arguably, much success that is only apparent (natural language inference) and mostly due to the very large datasets that are now available. Also, arguably, a lot of resources have been spent on something that seems more like a regression (deep learning chatbots, where it took a long time to rediscover basic notions like coherence).

Is dialogue modelling perhaps not a task that is usefully approached end-to-end? There are many interesting questions here about how to approach language science and engineering in general. (E.g., relation btw. machine learning and human learning; modularity; ... I've touched upon some of these in a recent manuscript on ArXiv, "language tasks and language games", [5].)

Is building conversational agents AI complete?

Unrestricted conversation clearly is AI complete (ie., requires the full realisation of intelligence): Any problem-solving behaviour of a person can be "simulate" in conversation (by imagining and describing it), so if any behaviour of humans is intelligent, conversation is also it. But conversation is also more narrowly and directly very challenging, not just because of what can be talked about, as it involves planning, reasoning, and acting under time-pressure.

The question is whether restricted conversation is not AI complete, and whether restricted conversation exists. Allen et al. [1] assume that it is, and does: "The Practical Dialogue Hypothesis: The conversational competence required for practical dialogues, while still complex, is significantly simpler to achieve than general human conversational competence" (see also [6, 3]).

The jury is still out on this, I would say. When things work, the illusion can be created that conversation is happening. When things don't anymore, this illusion breaks down very quickly. It seems that a) recognising misunderstandings, even in restricted tasks / domains and b) recovering from them might be AI complete.

In "The Symbolic Species", Terrence Deacon [2] makes the startling observation that there are no simple languages. But maybe there are simpler language users (human first language learners), or simpler language addressees? (See References.)

Social agents that don't use language

There is a model organism for social interaction, though, which is human/dog interaction. There is a strong impression that some form of communication is happening. What are its limitations? Can you achieve reference with a non-linguistic agent, or only joint attention? What are its mechanisms? Modelling this won't tell us about the role of language in this, but it might tell us about the role of paralinguistic signals and of interaction management (monitoring, turn-taking).

(Exploring this idea with a roomba-type robot has been on my "grant proposal ideas" list for a long time.)

Language-using agents that aren't social

Is it possible to build language interfaces that avoid giving the impression that they are more capable than they are? Interfaces that don't say "I", and don't pretend to be an "I"? Is that a useful goal?

It seems to me that a lot of the frustration that users have comes from them expecting "normal" language behaviour from systems, which they can only provide within a very narrow corridor of choices. Maybe that is a problem that goes away with exposure (perhaps Siri, Alexa, etc. have by now trained their users better?). Or is there space for designing interfaces that avoid giving the impression of having capabilities that aren't really there? It works for command line interfaces, but these are only for experts.

In [4], we tried to explore how using a non-speech modality could keep some aspects of conversational behaviour (quick feedback; prediction; adaptation / common-ground) while otherwise exposing limitations (restricted set of expectations). Many more design cues could be explored (e.g., "robotic" voices).

References

- 1 James F. Allen, Donna Byron, M. Dzikovska, George Ferguson, L. Galescu, and A. Stent. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3), 2000.
- 2 Terrence Deacon. *The Symbolic Species*. Norton & Co, 1997.
- 3 Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. Towards human-line spoken dialogue systems. *Speech Communication*, 50:630–645, 2008.
- 4 Casey Kennington and David Schlangen. Supporting Spoken Assistant Systems with a Graphical User Interface that Signals Incremental Understanding and Prediction State. In *Proceedings of the 17th Annual SIGdial Meeting on Discourse and Dialogue*, 2016. <http://clp.ling.uni-potsdam.de/publications/Kennington-2016.pdf>
- 5 David Schlangen. Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research; *CoRR 2019*, 2019. <http://clp.ling.uni-potsdam.de/publications/Schlangen-2019-1.pdf>
- 6 David Schlangen. What we can learn from Dialogue Systems that don't work: On Dialogue Systems as Cognitive Models. In *Proceedings of DiaHolmia, the 13th International Workshop on the Semantics and Pragmatics of Dialogue*, SEMDIAL 2009, pp. 51–58, Stockholm, Sweden, 2009.

5.20 Interaction Model for SLIVAR

Abhishek Shrivastava (Indian Institute of Technology – Guwahati, IN)

License © Creative Commons BY 3.0 Unported license
© Abhishek Shrivastava

Interaction models or Conceptual models are well defined in the area of Human-Computer Interaction (HCI). To quote Preece et al. ([2], p. 40), Interaction model is

a description of the proposed system in terms of a set of integrated ideas and concepts about what it should do, behave and look like, that will be understandable by the users in the manner intended.


These models are evaluated across three dimensions: (a) descriptive, (b) generative and (c) evaluative [1]. This means that an interaction model can help designers describe a range of possible interactions between the human and the computer, generate newer interactions within a specific conceptual framework, and (c) evaluate interactions across a range of design alternatives. I argue that a similar conceptual understanding of interactions between the humans and the virtual agents and robots is yet to substantiate. The research and design community are still required to arrive at a common understanding of conceptual models which drive the design and developments of SLIVAR. Not only that, we need to evolve a common understanding of these models, we need to find relevant candidates which can be evaluated across descriptive, generative and evaluative dimensions. During SLIVAR Dagstuhl seminar, it was seen that the community was hugely concerned about articulating these models. Anthropomorphism and, at times, animal-like did come across as existing conceptual models. However, there were open questions on evaluating interactions born out of these models. I suspect that unless we find methods to evaluate these interactions, we may be designing only limited scope point-designs. In my proposal, here, I believe that the answer lies in finding relevant interaction models for SLIVAR. This may be an interdisciplinary research where we may as well utilise methodologies involved in evolving Interactions models (as in HCI).

References

- 1 M. Beaudouin-Lafon. Designing interaction, not interfaces. In *Proceedings of the working conference on Advanced visual interfaces – AVI '04*, p. 15, New York, New York, USA, 2004. ACM Press. <https://doi.org/10.1145/989863.989865>
- 2 J. Preece, Y. Rogers and H. Sharp. Interaction Design: Beyond Human-Computer Interaction. *Design*, 18, 2007. [https://doi.org/10.1016/S0010-4485\(86\)80021-5](https://doi.org/10.1016/S0010-4485(86)80021-5)

5.21 Personal Statement on Spoken Language Interaction with Virtual Agents and Robots

Gabriel Skantze (KTH Royal Institute of Technology – Stockholm, SE)

License  Creative Commons BY 3.0 Unported license
© Gabriel Skantze

Long-term benefits of human-robot interaction

Social robots are often perceived as more engaging than other speech interfaces (if we may call them that), such as avatars or smart speakers. To what extent is this only a novelty factor? What happens when this wears off? What are the long-lasting benefits of human-robot interaction, compared to other forms of spoken interaction? Since robots are much more expensive, the advantages they provide must be very clear. In theory, and intuitively, it is clear what these advantages are. I often make the analogy that we are very reluctant to have important meetings (and would never take a Friday beer) over Skype or over telephone. So, there is clearly something special about physical face-to-face meetings and situated interaction (even if we do not actually interact physically). But the scientific evidence for the benefits of human-robot interaction are not abundant. Several studies have shown benefits of robots in for example educational settings [1]. But recent large-scale studies have not found these effects [4]. How can we go about understanding these phenomena better?

Anthropomorphism

Given that we think that social robots provide an added value compared to other speech interfaces, another important question is whether they should look like humans? Most social robots of today (NAO, Pepper, etc) are not very human-like, and their faces are not very expressive. Other robots (such as Jibo) are certainly expressive, but not human-like. Two of the most common arguments against human-likeness are: (1) the Uncanny Valley, and (2) the increased expectations from the user (which cannot be met). I would argue that (1) is certainly a problem for robots like Sophia, and is essentially a problem of mismatch between behaviour and appearance. However, we do not typically find human-like animated agents (like those found in Pixar movies) uncanny, so there is no reason why that would have to be the case. Regarding (2), I think robots situated in a specific setting (such as a reception) can help to limit the expectations (you would not ask a human receptionist for the meaning of life). I think that one of the strongest argument for human-likeness is that the face helps to coordinate the interaction, as it carries a large set of social signals that we already know how to interpret and can relate to [2].

Deep learning for conversational agents

While there have been enormous benefits of deep learning in other areas of speech and language processing (ASR, TTS, MT, etc.), this is not really the case for dialogue systems. I think there are at least four explanations for this: (1) The mapping between input and output is much more indirect for a dialogue system as a whole, compared to ASR, TTS, MT, etc. Simply put, there are many potential answers to the same user utterance or dialogue context. Standard measures like BLEU, etc., used for other technologies do not apply very well. (2) Dialogue is inherently interactive, which makes dialogue systems inadequate to evaluate using fixed datasets. Thus, the ultimate test for a dialogue system is interactive challenges such as the Alexa challenge, but these are very expensive to perform (and hard to define in a meaningful way). (3) Dialogue systems operate in specific domains where there is often not much data to train on. (4) End-to-end training of dialogue systems goes against the idea of modularisation. Thus, it is not clear how such a system could be updated with new vocabulary, database items, etc., without retraining the whole system and complement the dataset with new examples used in the specific contexts, etc. So, the question is how we can make use of deep learning for dialogue systems (beyond the individual components)? Personally, I think representation and transfer learning for dialogue is an interesting topic that should be investigated more. At KTH, we are currently looking into how models of turn-taking can be learned from human-human data and transferred to human-computer dialogue, using deep learning [3].

References

- 1 D. Leyzberg, S. Spaulding, M. Toneva & B. Scassellati. The Physical Presence of a Robot Tutor Increases Cognitive Learning Gains. In *34th Annual Conference of the Cognitive Science Society*, 2012.
- 2 G. Skantze. Real-Time Coordination in Human-Robot Interaction Using Face and Voice. *AI Magazine*, 37(4):19, 2016.
- 3 G. Skantze. Towards a General, Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks. In *Proceedings of SIGdial*, 220–230, 2017.
- 4 P. Vogt, R. Van den Berghe, M. de Haas, L. Hoffman, J. Kanero, E. Mamus et al. Second Language Tutoring using Social Robots: A Large-Scale Study. In *Proceedings of HRI*, 2019.

5.22 SLIVAR and Language Learning

Lucy Skidmore (University of Sheffield, GB)

License  Creative Commons BY 3.0 Unported license
© Lucy Skidmore

The study of SLIVAR in relation to second language acquisition is a rich topic which has been explored at length in the field of computer-assisted language learning (CALL), continuously diversifying as new technology emerges. The evolution of speech technology in particular has created new ground for exploration in dialogue-based computer-assisted language applications. With this new territory comes challenges and opportunities, some of which are highlighted below.

Technical challenges

Despite vast improvements in accuracy of automatic speech recognition (ASR) for non-native language, it can still fall short when used in language learning applications. Accommodating these scenarios is a fundamental challenge for researchers in CALL and imaginative steps need to be taken to both minimise the occurrence of errors but also navigate inaccurate recognition in a constructive way for learners.

Choice of platform

With the continuous accommodation of new technologies comes the vast array of platform choice for language learning applications. Naturally accessibility plays an important role in this decision – both robot-assisted language learning and mobile-assisted language learning are well-established sub-fields within CALL, however mobile-assisted language learning research has more direct impact on learners due to the accessibility of smartphones. With the increasing ownership of products such as Alexa and Google Assistant, voice-assisted smart devices are more accessible compared to robots. This is one of the important factors to consider when making a choice between platforms.

Role of anthropomorphism

The fact that this research is concerned with non-native speaker-computer dialogue rather than native speaker-computer dialogue raises interesting questions about the role of anthropomorphism in human-computer dialogue for language learning. How important is it for learners to hear human speech? Is synthesised speech sufficient for language practice? Are virtual agents and robots appropriate communication partners for learners?

What learners want


Motivation to learn has been shown to be a strongly influential factor in successful language acquisition. It is therefore essential for any research in this area to take learners' opinions into account in order to understand how learners want to use these systems. This in turn may provide interesting opportunities for applications not previously imagined.

An interdisciplinary approach is key

Interdisciplinary in its nature, any successful research into the applicability of SLIVAR to language learning will face the challenge of understanding the topic from multiple perspectives. These include but are not limited to second language acquisition, speech technology (ASR, speech synthesis and dialogue modelling), human-robot interaction and psycholinguistics. However, with this challenge comes the chance for collaboration amongst research communities, which is perhaps the most exciting opportunity of all.

5.23 SLIVAR and the Role of the Body

Serge Thill (Radboud University Nijmegen, NL)

License  Creative Commons BY 3.0 Unported license
© Serge Thill

Human signal interpretation is multimodal

Embodied cognition has long pushed the idea that sensorimotor areas of the brain are involved much more in all aspects of cognition than classic theories allow for. Relatedly, it has also become understood that human perception is multimodal, and thus not just focused on one sense at a time.

In particular, studies have demonstrated that humans are very attuned to perceiving biological motion even in the context of HRI: when a robot moves with biological kinematic profiles, humans imitate both the action and the speed at which the robot moves, but if the robot uses other types of kinematics, humans are no longer sensitive to the speed of the action, imitating only the goal [2]. This demonstrates that information from the visual modality may be augmented by motor information, provided that the observation can be parsed in human motor terms.

The first question is therefore whether similar processes might apply to sound processing: are we better at understanding speech when the speech is comprised of sounds for which we have a motor programme? If so, can HRI profit, and if so, how? For example, would a detailed morphological model of human sound production help a robot and are there robot vocalisations that are easier to understand for humans because they map onto human motor programs?

Human language is grounded

Genuine vocal interaction requires, in particular on the part of the machine, an understanding of the meaning of the concepts used. From a computational linguistics perspective, it has been clear for a while that purely statistical approaches on the textual modality is insufficient [5]. While this has been debated for quite a while already, the core question remains: to what degree would a robot need to “understand” the sensorimotor experience underlying human language, what exactly are the mechanisms of grounding, and does the fact that humans and robots have different bodies impose any limitations on the degree we can communicate [6]?

Do we even need sophisticated vocal interaction?

Most, if not all of present-day HRI operates in relatively constrained scenario and we are pretty far away from idealised generic “robot companions”. This raises the question whether there is a genuine added benefit to providing vocal interactions in realistic scenarios. Even in situations where robots are specifically meant as companions, for example as companion robots for the elderly, there is evidence to suggest that end users are satisfied with animal-like command interactions [3].


An interesting question to explore is therefore where exactly the added benefits of vocal interaction lie. It is clear, for example, that sophisticated vocal interaction may reduce the need to interpret other types of social signals, which is also a non-trivial problem [1] that taps into embodiment and would appear to require advanced models of Theory of Mind [4]. But what shape does this trade off take exactly?

References

- 1 M. E. Bartlett, C. E. R. Edmunds, T. Belpaeme, S. Thill and S. Lemaignan. What can you see? identifying cues on internal states from the movements of natural social interactions. *Frontiers in Robotics and AI*, 6:49, 2019.
- 2 A. Bisio, A. Sciutti, F. Nori, G. Metta, L. Fadiga, G. Sandini and T. Pozzo. Motor contagion during human-human and human-robot interaction. *PLOS ONE*, 9(8):1–10, 2014.
- 3 H. L. Bradwell, K. J. Edwards, R. Winnington, S. Thill, and R. B. Jones. Companion robots for older people: importance of user-centred design demonstrated through observations and focus groups comparing preferences of older people and roboticists in South West England. *BMJ Open*, 9, 2019.
- 4 H. Svensson and S. Thill. Beyond bodily anticipation: internal simulations in social interaction. *Cognitive Systems Research*, 40:161–171, 2016.
- 5 S. Thill, S. Padó and T. Ziemke. On the importance of a rich embodiment in the grounding of concepts: perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, 6(3):545–558, 2014.
- 6 S. Thill and K. Twomey. What’s on the inside counts: A grounded account of concept acquisition and development. *Frontiers in Psychology: Cognition*, 7:402, 2016.

5.24 What should an agent’s identity be?

David R. Traum (USC – Playa Vista, US)

License  Creative Commons BY 3.0 Unported license
© David R. Traum

In our conceptual structure, mirrored also in grammatical categories in many languages, we have different types of entities we consider and engage with. Agents, exemplified by people, have cognitive structure (e.g. beliefs, desires, intentions, emotions) and can be the moral and physical causer of actions. Objects can be acted upon (moved, modified, created, destroyed), and tools or instruments can be used by an agent to facilitate an action. Instruments thus can play a role in action, but usually without the conceptual structure and with another causer (agent) as taking ultimate responsibility. Interfaces and even human languages themselves could be seen as a kind of tool that allows humans to communicate with each other and with the physical and virtual worlds. When we come to robots and “virtual agents”, the question immediately arises as to whether they better fit the agent or instrument categories? Despite the name “agent”, many researchers and users take the instrument view – that the entity is there primarily to serve a human user or enhance their abilities, by allowing them to focus on a higher level of problem. Taking this point of view, the main goal of interaction should be efficient task completion, with a minimum of time or cognitive overload spent. On the other hand, if the agent perspective is taken, we would anticipate more effort spent at social-relationship building, empathy and perspective-taking and more symmetric interactions. While the tool view seems too limiting for many of the uses people desire to put robots and virtual agents, few, if any, are comfortable seeing these artificial entities as fully competent humans. While we don’t have many examples of intermediate types of entities, we do have some. Examples of entities that are seen as having some cognitive and moral agentive capacity but not fully competent members of human society include animals, children, and at least for some foreigners (who don’t fully grasp the language or culture), slaves, mentally ill or incapacitated, and criminals. While these roles vary from culture to culture (and culture views change across time), what is common is that members of these classes are generally

seen as having some of the conceptual structure and moral responsibility of full persons, but not all of it. These patterns may be a good launching point for how to construct and think about robot and virtual agent identity, as they often have been in fiction. A problem is that we often have conflicting intuitions about the source of rights to person and agenthood, e.g. whether it is based on biological relatedness or physical and cognitive abilities, or potential for these. There is often also a disconnect between how an entity portrays itself, how it is perceived, and the actual abilities. We are often willing to take quite superficial displays as signals of a host of abilities and attitudes, whether these displays come from other people or other types of entities. There is thus a potential for deception, which might be either benign, neutral or harmful to the specific interaction or for trust and expectations of future interactions. My position is that the identity and supporting displays for an artificial agent should match the role it is expected to play in desired interactions but also its capabilities as required by that interaction. Human-like activities require human-like identities. Some capabilities could be assumed while others must be demonstrated. Key also will be strategies for maintaining identity as well as the fluidity of interaction across communication errors and problems, especially as these can be seen as opportunities for building social relationships rather than just places where the system seems to be malfunctioning.

5.25 Are we building thinking machines or are we illusionists?

Preben Wik (Furhat Robotics – Stockholm, SE)

License © Creative Commons BY 3.0 Unported license
© Preben Wik

When we talk about using Spoken Language to Interact with Virtual Agents or Robots it is a very intuitive thing for the non-expert to understand. Because people are so adept at conversing with each other, it is fairly easy to understand how it should be working on a high level of abstraction. A lot easier than it is to actually implement it. Lots of really smart people are working really hard on it, yet a 5 years old kid will often do better than today's state-of-the-art conversational systems. People wonder what is so hard?

Of course I don't have a solution or quick fix for how to make it work well, but I am hoping that we will spend some time together these days to look at some of the bigger, structural challenges from new angles, and ask ourselves some new questions, and that way perhaps come up with some fresh ideas.

We may ask ourselves: What is it that the kid is so much better at? How come we are not able to do that aspect well with machines yet? Do we even know what these difficulties and challenges are? Let's consider the possibility that we might be barking up the wrong tree.

Most AI systems today are doing "Neo-cortex kind of stuff". Can we take a closer look at what cognitive tasks are done in the reptilian brain, or the limbic system and how they relate to our task? Perhaps there are some lower level features needed to make a system more responsive, and feel more "alive"?

I have worked in the chatbot industry, and the social robotics industry, and I have built systems for language learning (CALL), and for human-animal interaction. Although they are all about spoken interaction between different agents, I find it interesting that the questions asked in the various contexts are so different from each other.

While working with Human-Dolphin interaction some linguists say “They Can’t do it because animals don’t have a language acquisition device -LAD!” That statement is never heard in a SLIVAR context. We have had long academic debates about the origin of language and the source of our ability. Is it a divine quality? Do we have a language instinct? Do we have a LAD in our brains? If so, where does it reside? And what exactly does it do? Could we make an artificial LAD?

What should the building blocks for creating conversational AI be? Could there be something missing in our toolbox? Human-robot interactions are today typically built by writing scripts with some tools such as FurhatSDK, Watson, DialogFlow, LOUIS, Chatscript, Teneo etc. using building blocks such as Entities, Intents, and Topics. We need units on several different levels of abstraction. If we look at “a body” as an analogy, we can talk about it on different levels: atoms, molecules, amino acids, cells, joints, tendons, muscles, organs, arms legs etc. But we cannot understand a kidney from a molecular-level. It will just be a big mess of a bunch of molecules. Similarly, how can we deal with implicatures in a “Gricean” sense for example? Are we able to capture the cooperative principles described in his “logic and conversation”? What about replicators such as “memes”? And what about metaphors and analogies?

As we stand now, we should not forget that we are in the illusion business. Today our job is to create the illusion that you are talking with something that understands you and cares about what you are saying. Which it doesn’t. There is nothing wrong with that, but it is a distinctly different discipline from the engineering of building thinking machines. Is that where we see ourselves heading? Is there another path where we are building sentient machines?

Participants

- Hugues Ali Mehenni
CNRS – Orsay, FR
- Gérard Bailly
University Grenoble Alpes, FR
- Bruce Balentine
Entreprise Integration Group –
Zürich, CH
- Roberto Basili
University of Rome “Tor
Vergata”, IT
- Timo Baumann
Universität Hamburg, DE
- Michael C. Brady
American University of Central
Asia, KG
- Hendrik Buschmeier
Universität Bielefeld, DE
- Nick Campbell
Trinity College Dublin, IE
- Nigel Crook
Oxford Brookes University, GB
- Laurence Devillers
CNRS – Orsay, FR
- Johanna Dobbriner
TU Dublin, IE
- Jens Edlund
KTH Royal Institute of
Technology – Stockholm, SE
- Mary Ellen Foster
University of Glasgow, GB
- Emer Gilmartin
ADAPT Centre – Dublin, IE
- Manuel Giuliani
University of the West of
England – Bristol, GB
- Martin Heckmann
Honda Research Institute Europe
GmbH, DE
- Kristiina Jokinen
AIST – Tokyo Waterfront, JP
- Tatsuya Kawahara
Kyoto University, JP
- Casey Kennington
Boise State University, US
- Evangelia Kordoni
HU Berlin, DE
- Ivana Kruijff-Korbayová
DFKI – Saarbrücken, DE
- Pierre Lison
Norwegian Computing
Center, NO
- Joseph J. Mariani
CNRS – Orsay, FR
- Cynthia Matuszek
University of Maryland,
Baltimore County, US
- Roger K. Moore
University of Sheffield, GB
- Mikio Nakano
Honda Research Institute Japan –
Wako, JP
- Catherine Pelachaud
Sorbonne University – Paris, FR
- Roberto Pieraccini
Google Switzerland – Zürich, CH
- Matthias Scheutz
Tufts University – Medford, US
- David Schlangen
Universität Potsdam, DE
- Abhishek Shrivastava
Indian Institute of Technology –
Guwahati, IN
- Gabriel Skantze
KTH Royal Institute of
Technology – Stockholm, SE
- Lucy Skidmore
University of Sheffield, GB
- Serge Thill
Radboud University
Nijmegen, NL
- David R. Traum
USC – Playa Vista, US
- Matthew Walter
TTIC – Chicago, US
- Lun Wang
Sapienza University of Rome, IT
- Preben Wik
Furhat Robotics – Stockholm, SE

