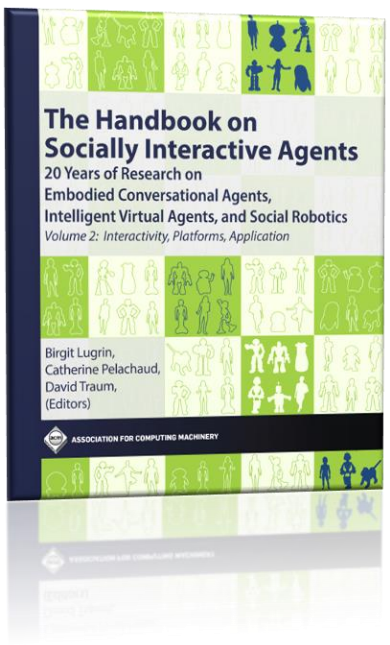
# Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics

Birgit Lugrin, Catherine Pelachaud, Elisabeth André, Ruth Aylett, Timothy Bickmore, Cynthia Breazeal, Joost Broekens, Kerstin Dautenhahn, Jonathan Gratch, Stefan Kopp, Jacqueline Nadel, Ana Paiva and Agnieszka Wykowska




**Author note:**

This is a preprint. The final article is published in "The Handbook on Socially Interactive Agents" by ACM.

**Citation information:**

B. Lugrin, C. Pelachaud, E. André, R. Aylett, T. Bickmore, C. Breazeal, J. Broekens, K. Dautenhahn, J. Gratch, S. Kopp, J. Nadel, A. Paiva and A. Wykowska (2022). Challenge Discussion on Socially Interactive Agents: Considerations on Social Interaction, Computational Architectures, Evaluation, and Ethics. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *The Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 2: Interactivity, Platforms, Application (pp. 561-626). ACM.

DOI of the final chapter: 10.1145/3563659.3563677

DOI of volume 2 of the handbook: 10.1145/3563659

Correspondence concerning this article should be addressed to Birgit Lugrin, birgit.lugrin@gmail.com, and Catherine Pelachaud, catherine.pelachaud@isir.upmc.fr

# A Challenge Discussion on Socially Interactive Agents: considerations on social interaction, computational architectures, evaluation, and ethics

Birgit Lugrin*, Catherine Pelachaud*, Elisabeth André, Ruth Aylett, Timothy Bickmore, Cynthia Breazeal, Joost Broekens, Kerstin Dautenhahn, Jonathan Gratch, Stefan Kopp, Jacqueline Nadel, Ana Paiva, Agnieszka Wykowska

* Birgit Lugrin and Catherine Pelachaud have a shared first-authorship on this chapter.

## A.1 Introduction

This chapter contains a collection of interviews on current challenges and future directions that researchers are faced with when working with Socially Interactive Agents (SIAs), see chapter 1 on "Introduction to Socially Interactive Agents" [Lugrin 2021] of volume 1 of this handbook [Lugrin et al. 2021] for a definition. The works reported in all the chapters of Volume I and Volume II of this book have highlighted the importance and necessity to take an interdisciplinary approach when conducting research on and developing SIAs. It requires dealing with many facets of multimodal behaviours that occur during an interaction between humans and other agents that can take place in a great variety of social domains. When working on finding key challenges that still need to be addressed, different clusters of questions were built and several experts in their respective fields were invited for the interviews. They thus discussed various aspects of the research with SIAs from different perspectives and laid ground for lots of future research directions, introduced thought-provoking ideas, and discussed the potential risks of this research area.

The interviews were conducted by two of the editors of this handbook:

**Birgit Lugrin**: Professor for Media Informatics at the Julius-Maximilians-University of

Würzburg, Germany, and
**Catherine Pelachaud**: Director of Research at CNRS in the laboratory ISIR, Sorbonne University, France,
while the interviewees were authors of various chapters of both volumes of this handbook.

We prepared the interviews ahead of time by following a bottom-up methodology. When we outlined each of the chapters in the very beginning of this handbook, we asked every author to include a section on current challenges and future directions within their specific research domain of SIAs. We first ran through the chapters and gathered the challenges they addressed. Not surprisingly, there were several overlaps that faced similar issues or risks that were of importance for various implementations, but with a different focus. They covered very broad issues, addressing the need of novel computational approaches, evaluation protocols, but also societal and ethical issues. Then, as a next step, we defined a set of main topic areas and clustered them into four main topics:

- Social Interaction
- Computational Architecture
- Evaluation
- Ethics

For each of these topics, we defined a set of open questions to be addressed during the interviews. The questions addressed SIAs in both their potential embodiments, as Intelligent Virtual Agents (IVAs) or Social Robots (SRs). We planned for four interviews, one on each of the defined topic areas, and organized interviews with two or three experts, who were all authors of different chapters of this handbook. A fifth interview dedicated on ethics in the application of SIA for children with Autistic Spectrum Disorders (ASD) was also organized with a specialist in this area and that has also written a chapter.

We told all interviewees that the interviews will be conducted in a semi-structured manner, via video conferencing. They were informed that the interviews will be recorded and transcribed. The transcription was done semi-automatically, relying on automatic transcription tools, but manually going through the whole interviews and making corrections afterwards. Then, the draft of the transcription of each interview was sent to all the interviewees of the respective topic area for potential corrections and final approval.

This chapter is organized as follows: each identified topic area (and the subtopic on ethics in ASD research) is a section of this chapter, introducing the interviewees as well as the list of questions that were addressed, and then reporting the interviews (sections 2-6). At the end of the chapter, some concluding remarks are given (section 7).

## A.2 Interview 1: Social Interaction

For our first interview, taking place in October 2021, we have drafted the following five questions:

Question 1: How shall we integrate social functions to facilitate adaptation, rapport, or engagement into the interaction with SIAs?

Question 2: How will we consider individualization of the SIA to match different personalities, genders or cultures?

Question 3: How much formality and natural language, e.g. politeness, do we need? Should we have to say, for example, "Alexa play Netflix please" or simply give a command?

Question 4: And from the agent's point of view? How much formality is needed here?

Question 5: Are robots the new IVAs? How do you foresee the potential of augmented reality?

### A.2.1 Participants

**Cynthia Breazeal**, Professor and Associate Director at the MIT Media Lab and the founding Director the Personal Robots Group

**Jonathan (Jon) Gratch**, Research Full Professor of Computer Science and Psychology at the University of Southern California and Director for Virtual Human Research at USC's Institute for Creative Technologies

**Ana Paiva**, Professor of Computer Science at INESC-ID, Instituto Superior Técnico, University of Lisbon

### A.2.2 Question 1:

Catherine Pelachaud: And the first question we had was, how should we integrate social functions to facilitate adaptation, rapport, or engagement into the interaction with socially interactive agents?

Ana Paiva: I would start with one thing related with the development of social agents. Nowadays, we have a lot of discussions about whether to have a theoretical driven approach to develop our agents' social behaviors, or a more "data-centred approach". I think we need to consider that the two can go hand in hand, especially now with all the machine learning techniques that supports these very data-centred approaches. But those data-centered approaches are not enough, and I believe that without theories, like theories of emotion, theories about report, that support building models at a more abstract level, we may be re-inventing the wheel. So I believe that the one of the big challenges we have now is to combine these two approaches for development and embrace both the data driven in certain aspects, as well as the theoretically driven approaches, as more symbolic AI people used to do. But I am not sure whether Cynthia and Jon agree with me...

Jon Gratch: Historically, I have been a strong proponent of theoretical approaches. I'm somewhat changing my thinking on that in a sense. And I guess one of the issues, at least within affective computing, is in some ways the prominent theories have led the field astray in the sense that there's been a predominant focus in affective computing on trying to recognize people's feelings. And that comes from work by Paul Ekman[1] who had a very strong sway over the field. Yet, it's not clear that that approach is actually supported by the data. Ekman's influence is much reduced in psychology, yet if you look at what companies are actually doing in the field they're tied to his theory. And that can sometimes lead things astray. Sometimes the theory shapes the questions we ask, and sometimes those aren't the right questions. So, I'm struggling with how to do that and I definitely believe we know that there's a lot of problems with data driven, and you get these black boxes and they learn something stupid when you finally figure out what they've learned. So we need to tie the learned models back to constructs that we know are actually tied to the phenomena we're trying to study. But somehow, some balance between the two is important and I guess that's what you're probably saying.

Ana Paiva: Yes.

Cynthia Breazeal: Yeah, and I agree. Both are important. Of course it begs the question of what data is informing what theories. I suspect we probably feel that we don't have the comprehensive kind of data sets that we would like to see across: geography, different groups, ages, cultures, etc. There is a broad diversity of people who we would like to have these systems to interact with, in richer and more competent ways. So, for my own work, we're looking more and more at longer term deployments, integrated into real human contexts, where people are going about their daily lives. We are trying to capture richer, more representative data that I think is going to challenge and force us to improve our theories. I see it as a virtuous cycle. A critical piece, I think, is challenging ourselves to get richer, more representative data sets. Of course, this gets us into the ethics discussion on how we design these systems responsibly? How to deal with these questions in light of design justice, and so forth. These are really important considerations for the field as we integrate our systems into society.

Jon Gratch: I just want to point out that I really appreciate some of the work Cynthia has been doing in terms of long term interaction. Because when you talk about phenomena like rapport and social function, there is a fair amount of work in our community, but typically those are "one shot" studies where participants never develop any kind of long term relationship with the technology. So they don't actually get to learn if there is really a function to those expressions. Designers make something look like it has a function and people treat it as so in a short-term experimental study, but then the designers don't

[1] https://www.paulekman.com/

actually get around to actually implementing that function. In that companies are actually deploying this technology in the real world for persistent interactions, I think that makes that question of what are the function of these behaviors much more important, because it can look cute or look like it has a sophisticated cognition, but if it doesn't follow through and actually do those things, then people just tune that kind of stuff out.

Ana Paiva: Yeah, I truly believe that it changes the way we look at the interaction and where the interaction is and so forth. When you look at one shot studies is different from when you look at long term interactions. In fact, we saw very clearly in an old study with the iCat[2] that when we had just one short interaction the novelty factor is there[3]. People look closely at the robots and our agents, as they are trying to figure out all the social signals, and enjoying the novelty of the situation. But then when the interaction follows on the next day, or next week, and becomes repetitive, a lot of the signals become irrelevant, and the task becomes more important. The salience of the social signals decreases. So I remember with the iCat that after several interactions, the kids didn't even look at the iCat anymore. They were just looking at the chessboard because that was the important they needed to look at, and the iCat, apart from some very specific moments, didn't matter anymore. So there is a clear change, especially for the social signals, when they interact over long periods of time. Also because they realized that the robots, or the agents, could not see them or cannot follow the same signals as we do, and don't respond the same way. So, in fact I remember at some point they would ignore what was going on because they knew that the system wasn't able to respond in a natural way as we humans respond. So, yes I totally agree.

Cynthia Breazeal: We are also finding that personalization is increasingly important as we delve into long term interactions. People are changing and the system as they interact with it, and ideally our systems will be able to adapt and change to continue to be engaging and helpful to people. So we have been looking at, for instance, agents in a family context where of course you have different kinds of people – different ages, different ways they interact – so personalization is becoming a more important theme for our work, as being informed from this long term interaction context.

### A.2.3 Question 2:

Birgit Lugrin: I agree, it is really important. And this really goes into our second question that is on personalization and individualization. So if you would like to elaborate a bit

[2] van Breemen, A., Yan, X., Meerbeek, B. (2005, July). iCat: an animated user-interface robot with personality. In Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems (pp. 143-144).

[3] Leite, I., Pereira, A., Martinho, C., Paiva, A. (2008, August). Are emotional robots more fun to play with?. In RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication (pp. 77-82). IEEE.

more on how we should consider individualization and how we should match different personalities, genders, cultures and so on.

Cynthia Breazeal: I say YES to all of that because people are not the same. Of course, personalization also brings ethical questions around privacy, transparency, and accountability. It is a rich, multi-faceted challenge beyond what algorithm we are going to apply to achieve longer term personalization, but we need to consider the broader context of what that means. Again, for people who are potentially living with these systems, I think it is also a really important consideration. I think cultural differences are important to capture, and we've already talked about different age groups, different applications. How the agent presents itself, and its role is going to vary depending on what the main kind of task or value proposition the agent is offering. In our work we're seeing things like differences in personality traits may influence the effectiveness of a given intervention for a given person. We found this in our emotional wellness work. We may discover this is another important aspect of personalization, which may be based on personality profiling – in addition to things like interaction style, and role, and all of these other things. It is a very rich research topic.

Ana Paiva: Yeah, I see that personalization is something that needs to be carefully considered, because it can actually make the interaction worse. I've seen some situations where the adaptation changes make the interaction not so fluid. So, if the agent is changing as it interacts with one person, if that personalisation is not well done, then we can have negative effects. And in fact, I wonder: How far do we want to go with personalization? Because given what's going on with social media, I really think that exaggerating the personalization and all the information that's captured about the individual user may have some ethical problems. We may need to think about personalisation in different ways. I did my Ph.D. many years ago on personalization and user modeling, and now, many years later, I'm questioning some of the fundamentals and what drives such personalization. Because I don't want to be sold something that the system already knows that I like. For example, I searched for trainers, bought some trainers, and now I keep getting trainers sold to me. I don't want that kind of personalization. I want the systems to know about the relevant things about users, that make the interaction more engaging. Or maybe engagement is perhaps not the right target, but rather that users can learn more. But knowing about my training shoes? I think we need to be very careful about personalization. And it's something that we as a community need to think about it. For example, with chatbots, do we want our chatbots to really get all kinds of information about us?

Jon Gratch: I think we need a better language to characterize different aspects. For a tutoring system, I think it's somewhat uncontroversial you want the tutor to personalize the feedback based on the particular errors the student has. You might think that for speech recognition

systems it's uncontroversial to train and tune the recognition on that particular person. Although, you can even take that to the extreme: so when the person starts diverging from normal English language, or starts doing racist things instead, do you want the agent to adapt and reinforce and support those kind of behaviors? I don't think so. But at a broad level, personalization is also an effective influence tactic. So companies, for example, want to allow people to customize and personalize, say their Alexa, or some other customer service app, so they feel it's their friend. But it's not their friend, right? It's the voice of the company. And it's there for a very specific purpose: to sell more stuff. More broadly, I'm somewhat conflicted about the idea, and Ana knows this because we were in a workshop together in Dagstuhl[4] about whether we should make these things. When you're talking about personalization you're using very anthropomorphic terms. So should we anthropomorphize these things and make them seem like humans, and seem like they care about us and have a relationship with us? Should human-human interaction be the gold standard, or should these somehow be different and maybe take advantage of the uniqueness of the technology to create other metaphors. There's a discussion at ACII[5], like is there such a thing as gender neutral speech? Because it seems to be that people want to personalize, or designers want to personalize their assistance to be women, right? Because that fits people's preconceptions, and then tends to reinforce those those cultural stereotypes. So I don't, and I don't have an answer there, but we need a better language to talk about these things.

Catherine Pelachaud: Yes, I'd like to add something. So when we talk about personalization, I don't think it means solely that we could, for example, choose the colors of the agent. It could be also understanding when to personalize and which factors to personalize. Is it possible? It means people would have to personalize the agent from the beginning to the end. I think it is this aspect that we should look at.

Jon Gratch: Even if you're the consumer and you're creating this agent for yourself, in some ways, do you really want to create an agent that sort of fulfills your wishes in every way, or do you want something that confronts you and challenges you instead? There's a sense in which we can easily make this technology narcissistic, right? It just reinforces a person's current foibles. That's why I think you're concerned, like Facebook is in a sense trying to do something like that.

Cynthia Breazeal: There is the question of how we approach the challenge of personalization and all of these different facets. And then there's also making sure that it supports what's

[4] Jonathan Gratch, Stacy Marsella, Arjan Egges, Anton Eliens, Katherine Isbister, Ana Paiva, Thomas Rist, Paul ten Hagen, "Design criteria, techniques and case studies for creating and evaluating interactive experiences for virtual humans", in Evaluating Embodied Conversational Agents, Zsofi Ruttkay, Elizabeth André, W. L. Johnson and Catherine Pelachaud (Eds.), Dagstuhl Seminar Proceedings 04121, 2006; https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=04121;

[5] International conference on Affective Computing and Intelligent Interaction - https://acii-conf.net/

important to people. Transparency, accountability, and explainability are going to be really important in addition to privacy and security. Privacy is often first and foremost in people's minds when they hear something like personalization. But, people need understand what it actually means for their system to personalize to them, and people should be able to further shape or change that. What if a system adapts to the person in a way that isn't right? People should be able to adjust it.

Ana Paiva: And not only that, I think the user's autonomy to decide whether they want to really get their data in the system, and the system to be adapting to themselves or not, it something important. The user needs to be in charge of that decision. And that's to do with transparency, but also guaranteeing the user's agency. That is something that I don't like, when my social media (Facebook or others) tell me to buy something, or recommend something that is assumed I would like- my agency is at stake. Plus, when I click "why am I seeing this advert", it reports to me some general justification like I'm a woman, I speak English, I live in Portugal, or I'm between 25 and 60 years old. And I know that this is not the reason why I'm seeing a certain advert. So, personalization must be linked with transparency in a way that guarantees user's agency.

### A.2.4 Question 3:

Catherine Pelachaud: We were thinking, and there's some work especially in social robotics, on how much formality and type of natural language do we need? Would we say "Alexa could you play Netflix, please", or simply give a command "Netflix, open"?

Jon Gratch: You could imagine Alexa, if you are not polite, would have a model of politeness and complain. Dialog and social norms get constructed through the interaction. You may have seen, when Manuela Veloso[6] gets on a soapbox about how technologists shouldn't encourage people to talk to robots like people. They are not people, and so why should we reinforce that with language? And I guess it just touches on the idea that do we leverage and reinforce existing stereotypes or not? There is some research around Alexa, specifically in children, where children learn certain interaction styles with Alexa and they learn that maybe it's fine to be bossy or rude. And then the question is, does that transfer, because it's human-like, to other humans? And I think it's unclear. It's the same problem with games, right? When people kill each other in games, does that make them more likely to kill each other in the real world? It is not clear. But I think it's hard to imagine language, without making it have to know something about how humans use language. And so I think at some level, we're sort of stuck with the fact that people have emotions, and they use those emotions to communicate things. They have social goals and those things probably have to be implemented and understood at some level by these machines, and then those machines need to reinforce some set of social interaction

[6] Manuela Veloso, Carnegie Mellon University, https://scholar.google.com/citations?user=2FbkAzYAAAAJhl

norms. What those choose to be, I am not sure. But I think eventually Alexa, it will try to enforce politeness norms.

Cynthia Breazeal: There's the natural language understanding, and then there is how that builds on this notion of what's the relationship? Which gets into roles, and appropriateness, and so forth. For example, when we designed Jibo[7], we gave a lot of thought into the way the robot speaks about itself. Jibo always reinforces "I'm a robot". And if you asked a particular question, like on religion, Jibo admits "I know nothing about that, you should ask another person". So through the agent itself, and how it contributes to the conversation, Jibo continually reinforces "this is what I am, this is what I can talk about, there's things I can't talk about, that you really just need to talk to other people about". Reminding and reinforcing Jibo's differences to being human was a part of our design philosophy. We are discovering as we go into longer term interactions, across age bands and application contexts, the one thing that people do want more from these agents – at least in the contexts we've explored like health, wellness, and education – people want more capable multi-turn conversations and dialogue with the agent. They want the agent, at least within appropriate bounded ways, to remember their conversations and past contexts. People want to avoid this constant repeating of what was said yesterday, or the day before, like the agent has short-term memory loss. At that point the agent just seems stupid, right? So, there's so many rich facets about how you do this well, in a way that respects people and their values, and what people want out of these systems. Drawing appropriate bounding boxes around these agents is really important. I think our field has been so caught up in the technical challenges, because of course there's so many technical challenges. But as these technologies are getting out there, and people are starting to interact with them, and expectations for what they can do and can't do are being established – these broader ethical design frameworks are becoming more and more important for us. We need to really develop new ethical design methods, and co-design methods. It's hard to say it's an all or nothing thing on any of these dimensions. It's about what's the right and responsibly bounded scope. And that just depends on a lot of things.

Ana Paiva: Yeah, I totally agree. Once you build your agent that is able to interact through natural language, it raises expectations from the point of view of the user, who expects the agent to learn and understand what is being said. So, natural language raises the expectations, and if the system does not meet those expectations there's a problem, as trust may decrease significantly. So, from a technical point of view it is hard, but on the other hand, if you don't do it, the interaction becomes limited, because people really want that type of interaction. So it's a balance that you have to juggle. Actually, in the

[7] https://jibo.com/

Dagstuhl event[8], that we were a couple of weeks ago, Catherine was there too, Roger Moore[9] was arguing that robots should speak with a robot voice. Robots, agents, or chatbots must disclose what they are, and talk with voices that are clearly mechanical, clearly not human so as not to raise too many expectations about the agent. So I think disclosing what it is, saying "I'm a robot", "I am a chatbot", and talking with the voice of a robot helps because then the natural language aspect becomes less demanding, and people don't feel cheated. I think managing expectations is important in natural language interaction between humans and robots.

Jon Gratch: But some other research shows that portraying yourself as a person is more effective for, for example, a mental health application. There's a number of studies where they actually manipulated robot backstory versus human backstory.

Ana Paiva: Are you saying that when the system pretends to be a human, it's more effective?

Jon Gratch: Yeah, so there are studies that show, it is more persuasive, but also elicits more honest disclosure. Somehow people find it more relatable. Even though in some sense of course they know it's a machine. Tim Bickmore[10] has done some of this work, and we have done some. So there's a tension between what we might perceive as unethical and what is effective. Hopefully there's not that tension, but sometimes there is.

### A.2.5 Question 4:

Birgit Lugrin: Thanks. That's already going in the next question, and I would really like to know what you all think about it, because that's the other way around: how much formality and politeness and things alike, are needed from the agent's side? Do you see a difference between what the user should be using for the interaction, and the agent should be using?

Jon Gratch: Well I think we are matching. If the machine doesn't adapt, people adapt to the machine. I think there's a natural tendency for synchrony, adaptation, and entrainment. I find I used a lot of these manners in terms of language with our virtual humans. People can't help but train yourself to the technology. So I think that choice will influence how people talk, at least with this genre of entities. It's a question of do you feel like it's important to reinforce certain social politeness and norms. I think people will change their behavior based on those two different choices, but as to what is the right choice, I am less clear about that.

Cynthia Breazeal: From my personal perspective, given what Jon was saying, we do see people mirroring the behavior of robots, and the attitudes conveyed by them. It is because we're

[8] Conversational Agent as Trustworthy Autonomous System (Trust-CA); http://www.dagstuhl.de/21381

[9] Roger Moore, The University of Sheffield, https://scholar.google.de/citations?user=Ib1I-uAAAAAJhl=deoi=ao

[10] Timothy Bickmore, Northeastern University, https://scholar.google.de/citations?hl=deuser=x9kzObUAAAAJ, author of chapter 24 on "Health-Related Applications of Socially Interactive Agents" [Bickmore 2022] of this volume of this handbook.

people, and we do these things when we interact with others. I think it's a skill, like any other skill. You reinforce the behavior you practice. So, if we want a society of people who are empathetic, compassionate, and polite, it behooves us to design systems that both convey and reinforce these behaviors, as well as to let people practice these behaviors. Because chances are, the behaviors that are reinforced over time become the default skills that people are going to apply to interactions across people, animals, and what not. So, let's look at it from a very human centered perspective: what do we want to encourage? I think it's important to consider the influences of these technologies, that are shaping us, both positive and negative. We have very transactional devices now, and we've seen examples of children being bossy or transactional with these devices.

Jon Gratch: It is less clear whether that generalizes to interactions with kids. We don't know.

Cynthia Breazeal: We don't know. Parenting plays a role, too. I'm just conjecturing that the more you behave in a certain way, the more that those behaviors are reinforced, the more it becomes default. Whether you intended to come across a certain way or not, I just see these as skills, practiced skills in general. I think these systems can play a role in helping us practice ways that we want to be as people, versus having us behave in ways that don't serve those goals. For me, that's more of a philosophical stance.

Ana Paiva: Yeah, I totally agree!

Cynthia Breazeal: We have data that shows that people do emulate the behaviors of these robots. I can imagine if people do that more and more, that will become more of a learned behavior. So, let us just be mindful of that when designing these systems.

Ana Paiva: I agree. My own chapter is about pro-social agents[11], promoting pro-social, and yeah I totally agree.

Jon Gratch: My presumption would be, the more that thing is like a human, the more likely it is to get generalization. The more the thing clearly indicates something different, it ought not to generalize. But the data from games at least is very unclear as to whether violence in games generalize this to violence in the school.

Ana Paiva: There are other aspects, such as promoting altruistic behavior. And, you actually can do that with games. There are systems that have shown that by interacting with a game they can help a lot to raise some more cooperative behaviour. Of course we don't know if it lasts over a long period of time. We've done one game for teenagers to help address the bystander problem in bullying. And what we found was, that by having this game and having the agents in the game, promoting these more altruistic and pro social behaviors in bullying situations. So I think that agents can be a force for good, and that needs to be

[11] see Chapter 11 on "Empathy and Prosociality in Social Agents" [Paiva et al. 2021] of volume 1 of this handbook [Lugrin et al. 2021].

further explored. I think there are avenues to explore and we as a community should be building these systems to promote a kinder society.

Cynthia Breazeal: To acknowledge this, it is something that we need to understand more - the social influence of agents on us, and these pro social opportunities. It's an area we need to understand more.

Jon Gratch: Yes, and it's interesting that there is a lot of work, and a lot of work here at ICT[12], where we are, and Ana does this as well, building agents to try to teach interpersonal skills, that we believe will generalize to the outside world. But then when we think about building a video game you say "that won't generalize the outside world". There is a difference in the sense that people know they're learning this skill, and they're encouraged to try to have the mindset to apply it in the outside world. Whereas in a game they are just having fun. But it does seem that this is a very blurry boundary, and if it's going to work in one, it's got to work in the other, right?

### A.2.6 Question 5:

Catherine Pelachaud: I think we are ready for our last question, we had foreseen. Which is, are robots the new IVAs, and how do you foresee the potential of augmented reality?

Cynthia Breazeal: I guess it depends how you define an IVA. The word "virtual" would make me think robots are physical, or physically embodied. So I wouldn't characterize any physical robot as a virtual agent. If we use the term "conversational agent", or something that's more agnostic to the embodiment, then robots can be considered emerging conversational agent that fits in with the smart speakers, and the animated agents, and everything that we already have. I think in the field of human-robot interaction, there is a growing interest in the expansion of more advanced multi-modal conversational abilities, in general. Virtual reality, augmented reality, other kinds of blended reality, and all different flavors of the amount of physical versus virtual permutations – we were interested in looking at things such as migratable AI. Right now, an AI persona might always be in a robot, or always in a particular device. But when you think about moving between contexts, from your home, to your car, to your workplace, to wherever: what if you have an assistive agent that can migrate across these embodiments to always be with you? I think is an interesting question. So, that rather than say it's a mixture, maybe it's also the transition between them. Migratable AI or migratable agents has been a research question for a number of years. And AI with all these conversational abilities is just bringing that to another stage in their development. But, I think any of these embodiments have their distinct set of advantages. A physical robot has the advantages that a physical body affords. Similarly, a virtual agent has the advantages that a virtual body affords. So it just

[12] Institute for Creative Technologies - https://ict.usc.edu/

really depends on your application in terms of what mix makes the most sense. It will just vary depending on the design context.

All: Yeah.

Ana Paiva: I think both virtual agents and social robots are embodied agents, that is, they have bodies. The robots are physically embodied in your environment, in our physical world, and are able to act there. The other ones are in the virtual world. Of course, like Cynthia said, there may be some blended realities. I've seen situations where you have a head mounted display and then your robot gets into the virtual world, so you can have all these combinations. But at the end of the day, embodied agents have properties that afford interaction through more senses, because you can see them, you can touch them (in the case of the robot), and they act in our world. So there's the body effect that I think is important in these embodied agents. And you can draw many conclusions from the virtual agents that can go into social robots, and the other way around. So the two communities should learn from each other, for sure, because there's many things that have been studied in the IVA community that the HRI community should have heard, and the other way around. So I think they are both conversational and "embodied" social agents, and that's it, and that's our field. In fact, I started working in social robots when I was doing virtual agents and went to one of our agent conferences, and I saw the iCat, and realised that I could use the same models that I had in my virtual agents in the iCat. So then, almost automatically, you go from one to the other. And I think that's the beauty of our field, the embodiment can be physical or can be virtual, and extend or combine these two.

Jon Gratch: My perspective is there's a thread within both robotic and virtual human community, that we want to build things that are like people. And that research is very similar between HRI and IVA if not completely overlapping probably. And it doesn't take advantage of the specificities of the modalities, and the only shame is that there's not more cross-talk. But as Cynthia was saying, there is uniqueness to the modalities, and in I think in robotic systems, we are actually forced to deal with the uniqueness of embodiment. It's about extreme power requirements, it's in a world that has to use actual sensors that we can cheat with the virtual human community. But the virtual human community seems more reluctant to explore what is unique or special or possible within the virtual and the augmented. And I think that's because they're too wedded to thinking of these things like real people. But you can transform the nature of the interaction, in augmented reality you can you see things that are not real, and as well as with virtual. I would encourage the virtual community and the augmented community to think more outside the human box and take advantage better of what unreality affords in terms of interaction to create something unique and special about the community. Otherwise, we're just to go to HRI.

Cynthia Breazeal: I actually think we, in the robotics community, we want to build human synergistic systems. I mean, it is a huge endeavor to build an android that actually looks and moves like a human, and most of the field is not concerned with that line of research. We're already in the design space of: robots don't look like people, they don't sense like people, but we want them to be compatible and able to interact with people, to be able to support people. So, for the robotics community, we don't view the human as the gold standard for how we want robots to appear and behave. Rather, we want to design robotic systems that are synergistic and complimentary to people. There are fascinating advantages to the differences between robots and people. How can we leverage this complementarity to enhance human-robot collaboration?

Ana Paiva: I believe that social robots should be designed in a way that the sociality of the robot has to come together with the tasks that it is going to do in the physical world. It's not enough just to be social, because then why do you need the robot? Why do you need it in the physical world? So I believe that exploring the physicality, changing things in your physical environment, and then on top of that adding the social interaction is one of the richness of the area of social robotics. Maybe that can be one of the differences from the virtual agents, I would say. But, at the end of the day, the two areas should go hand in hand.

## A.3 Interview 2: Computational Approaches

For our second interview, taking place in November 2021, we have drafted the following five questions:

Question 1: How shall we proceed to go beyond the SAIBA approach? Are incremental architectures the future?

Question 2: What is the next level of computational approaches? Is it machine learning only? Should we still consider symbolic AI and embed semantic information?

Question 3: How do we ensure transparency and explainable behavior?

Question 4: How shall we validate the computational architectures?

Question 5: Are robots the new IVAs? How do you foresee the potential of augmented reality?

### A.3.1 Participants

**Elisabeth André**, Full Professor of Computer Science and Founding Chair of Human-Centered Multimedia at Augsburg University.

**Joost Broekens**, Associate Professor of Affective Computing and Human Robot Interaction at the Leiden Institute of Advanced Computer Science (LIACS) of Leiden University.

**Stefan Kopp**, Professor of Computer Science and Head of the Social Cognitive Systems Group at Bielefeld University

### A.3.2 Question 1:

Catherine Pelachaud: The first question is : "how shall we proceed to go beyond the SAIBA approach? Are incremental architecture, the future?"

Stefan Kopp: This question presumes that we need to go beyond the SAIBA[13] approach and that there are any shortcomings or limitations with it, which is true. There are several ways you could go beyond it, and there're important things that should be worked on. One is to overcome its very coarse modular structure with three large modules. People have been proposing more fine-grained components, and nowadays we increasingly see machine learning approaches that try to do end-to-end learning trying to get rid of a modular structure. At the same time we still see more and more advanced approaches for highly specific modules. The other question is: are incremental architectures actually that important. I think absolutely yes, and we have been actually thinking on this right from the beginning. The SAIBA approach already includes the idea of incremental processing along its generation pipeline. It has BML chunking and we have been considering co-articulation effects, or feedback signals being sent back at different time scales. And I still think that incremental architectures are really important and it is something that

[13] Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson: Towards a Common Framework for Multimodal Generation: The Behavior Markup Language, International Conference on Intelligent Virtual Agents, 2006

we should embrace as a standard in the field. Especially if we want to build socially interactive agents that are responsive, that feel fluent and smooth to interact with, that are robust and also efficient when things get challenging, for instance in cases of communication problems.

Joost Broekens: I definitely agree with incremental and also modular approaches. I do feel that one element with this block-based processing process is that it seems to be still rather difficult to make the agents react fluidly on different time scales. So, fast reactive but meaningful reactions to stimuli and then some sort of a deliberative layer, I don't want to call it a subsumption architecture, but I get the feeling that it's still something that is needed. Let me put it this way. When I say this, it's biased by me working on social robots lately. I feel that especially in social robots, it seems to be, usually a rather monolithic approach, trying to develop for a particular use case. Then there doesn't seem to be much reactivity layers built in. I'm not sure if the SAIBA approach is very good at incorporating those kinds of fast reactive behaviors together with more high level behavior.

Stefan Kopp: The SAIBA pipeline that was originally spelled out is basically only a generation pipeline. So when we talk about reactive as opposed to more deliberative behavior, it's really about a full-blown architecture where we have to combine perception, processing, and generation via different routes, with different kinds of latencies, response times, but also depth of reasoning and planning ahead. Eventually we need incremental processing at all of these different layers of the architecture, at the respective time scales and being integrated with each other.

Elisabeth André: Essential features of incremental architectures are missing in the SAIBA architecture. For example, incremental systems operate on a more fine-grained time scale. Perceptions and responses co-occur, reducing latencies through parallel sensing, planning, and action. An essential requirement is smooth error handling. When we communicate with each other, we interrupt and correct each other. Interruptible output generation is a tricky business. An agent that plans too much in advance and cannot be stopped at any point may appear awkward. People seamlessly correct each other when needed. And unless we can simulate such behavior in an agent, the agent will appear unnatural. Another point is the continuous prediction of the conversational context of what we do. When someone starts talking, we try to predict what is going on. We might even jump in. We are constantly re-evaluating what we think about what might be next. It is worth thinking about which features of incremental architectures to integrate into the SAIBA framework.

Stefan Kopp: I fully agree. This is also a reason why I think incremental processing is crucial for socially interactive agents. Like Elisabeth was just saying, in social interaction it's

hard to predict what is going to happen next beyond a very short timeframe. It's also quite non-deterministic and thus complex to plan ahead. Classical planning approaches always hit the ceiling here in that regard. But incremental processing actually allows us to outsource this planning and processing to the interaction. By doing some limited, good-enough reasoning and then acting it out. Whether it works or not it's something that you will then see from how the interaction unfolds. Then you can respond to it. For example we produce utterances incrementally and process the feedback signals of the interaction partner to adapt our behavior online.

Catherine Pelachaud: It looks like you have mentioned three main profits about incrementality. One is for planning, one for managing interruption and another one is to handle behavior generation planning with machine learning that bypass behavior planner.

Elisabeth André: Also the error handling on the fly.

Catherine Pelachaud: They are four main features that are crucial to be handled with incremental processing.

Stefan Kopp: Although, for machine learning approaches it's also a challenge. These approaches often process more or less complete patterns and map them to some output based on correlation and features and so forth. If such patterns are not complete yet, as is often the case in incremental processing, this mapping is harder to learn.

Elisabeth André: Unless we use them for prediction, as is often done in dialogue generation. Machine learning approaches try to predict the next turn, the next word, or whatever. But of course, this approach comes with deficiencies on its own. It's not understanding. It's more predicting what to expect in a particular context. But do we want that for socially interactive agents? Maybe for some applications, we do not need a deeper understanding? But for other applications, we want to have agents with intentions and plans, agents who know what they are doing and know what the user is doing and not just show some seeming natural behavior.

Joost Broekens: At this last IVA conference there was a paper that I liked. They had an end to end trainable model. It was very clear that what was good in that approach is the fluidity of the movement. But if you would look at the iconic meaning of the gestures this was not that good. There is definitely a need, especially if you want to have full control over the communicative acts. You're going to go toward hybrid architectures with a symbolic as well as a machine learning approach. But the question is when do you do what and how do you mix them. Maybe you get the interpretation of the machine learning on the meaning of the gesture. But there is a lot to be said for both actually, shown by the striking difference between the communicative function and the fluidity.

### A.3.3 Question 2:

Birgit Lugrin: This discussion on machine learning actually already led us to the second question: looking into the future, what will be the next level of computational approaches? Is it machine learning only? We already know that it's probably not machine learning only. But what should be considered, for example, symbolic AI or embedded semantic information.

Elisabeth André: We will probably have a hybrid approach. Of course, it also depends on the application. It might be nice to have a chatbot just for fun and for your entertainment. But for some applications, the agent should have a deeper understanding of the conversation. I recently tested some chatbots. They gave the impression of a meaningful conversation, but only for a short time. After a few turns, they contradicted themselves and destroyed the illusion of an intelligent being. Also, most chatbots have tremendous problems with simple things like anaphora. The fundamental question is whether to focus on natural or intelligent behavior. We probably need both.

Stefan Kopp: I agree. I think it's key for agents to be adaptive in a very efficient way during the interaction, and to produce a behavior in a way that is responsive to the specific situation, to the specific interlocutor, to what was said before, to the unfolding discourse, and all these aspects. That actually requires an agent to be capable of really fast learning and fast adaptation. The hugely data intensive machine learning approaches struggle with this. I think the question here is not whether it's machine learning or not. It's whether it is statistical AI that is based on correlation patterns, generalized based on statistical significance, or it is model-based AI and machine learning where you extract structural knowledge regarding causes, effects and situation-specific parameters. I think we need both. A good example is gestures synthesis. We have models that produce body motion frame by frame, from acoustic or textual features. But they have no model of what they're talking or gesturing about, or what the bigger communicative context is. In result they can produce co-speech beat gestures very well, but not representational gestures. They lack fundamental levels of semantic and pragmatic aspects because they cannot be easily extracted from massive amounts of audio and video data.

Joost Broekens: You could imagine a system where, given that you have sufficiently rich labeled data, you could train the machine learning model to be able to cope with all kinds of variations of behavior. But you would be able to condition the model on the same sort of semantic information that model-based or agent-based approaches are able to get from their planning or from their reasoning engine. So then you can condition the machine learning model based on semantic information. I think that would be interesting. I would like to try it. It's a bit like speech, what Google and many other approaches are doing

for speech generation. You can do it with Tacotron[14], for example. You can actually condition the speech generation on quite a lot of features, if you can detect the features in the original data set. Then if you reverse the model, you can actually control those outcomes, and the speech that's produced by the model is using in essence symbolic conditioning information. Then you get emotionally varying speech. But it's not the model that figures out when to do it, not the machine learning model. It's an agent-based model that could figure out when to do it, to show empathy for example. So there are definitely interesting ways.

Stefan Kopp: It also leads us to the next question because model-based or symbolic AI techniques are needed to be interpretable or explainable.

Elisabeth André: The question, of course, is how much transparency we need in which situation. An agent that is supposed to induce a behavior change in someone may be more successful when communicating information subtly and indirectly. However, agents with implausible behavior cause problems in many situations. A few years ago, we developed an educational environment with virtual agents to help children deal with bullying at school. The children could advise the agents, and depending on the agents' emotional model, the agents either followed the children's advice or not. Some children got upset when the agents did not do what the children suggested. The children did not understand that the agents were afraid to carry out the recommended actions because the agents did not always convincingly demonstrate their fear. In this context, I would like to refer to an early IJCAI paper by Phoebe Sengers[15] on "Designing Comprehensible Agents." She argues that intelligibility should be an integral part of an agent's architecture rather than adding an explanation component post-hoc.

Stefan Kopp: But, then, the agent has to portray it in a way that is understandable and accessible to the human. It's obviously something different than explaining the 2 million parameters in a deep neural network. That's a common problem of explainable AI and it also applies to some of the models that we use in socially interactive agents. There's XAI[16] technologies to analyze black-box models and build interpretable models. But I think we have to bear in mind that we have two kinds of addressees, one is the developer and one is the user. XAI is really toward the developer, being able to analyze what the system is doing and what not and why. But we also have the user who needs to understand what the agent is doing and for what reason. And they need other explanations. The technology that we're developing actually may help to build better explainable systems because, in order to make themselves really understood, they must have abilities for dialogue or multimodal communication.

[14] https://google.github.io/tacotron/

[15] Phoebe Sengers: Designing Comprehensible Agents. IJCAI 1999: 1227-1232

[16] Explainable A.I.

Joost Broekens: When I think about transparency and explainability, I wonder why we want that so badly for this type of technology. I think there should be a discussion about why we want explainability in the first place for advanced technology. If you buy a car, I bet most people don't know how the car works, but it is predictable.

Stefan Kopp: No I don't agree. Laugh.

Joost Broekens: That's right for you. But a lot of people will only know how it works from an input output paradigm.

Stefan Kopp: Exactly.

Joost Broekens: If something goes wrong, if something unexpected happens, nobody knows why, then you go to a garage. As long your conversational agent behaves according to what you would expect, predict, in a particular setting, it is fine. But if at some point it says: "Well, maybe you should buy this medicine". That's weird. That's probably the moment where you would want it to be able to explain. But how do you do that. Explainable AI is almost always conversational.

Stefan Kopp: You would think so but in fact the field is still working to achieve quite simple forms of conversational explanations, for instance for recommender systems. We wouldn't even call it conversational explanation as the user is often only able to ask: "why are you recommending this hotel to me now" and then is presented with some additional information.

Elisabeth André: But it's not really a conversation. It is a follow-up question.

Stefan Kopp: Exactly, but it is called conversational explanation nevertheless. But what you're saying, Joost, is exactly right. Users have to have a good enough understanding of the system that is cognitively manageable for them and good enough for them to predict the behavior and make sense of it, and then it works fine.

Joost Broekens: And adaptive also, at least at some level; adaptive to what the user, at that point, needs. If you don't understand why someone says or does something then you ask the question: "Why do you do that", you get an answer at some level. Then you say: "well I still don't get this and this, can you explain", and you selectively go into what you need to know. That's very challenging.

Stefan Kopp: Yes, explainability has to fit to the user's information need.

Elisabeth André: There is an interesting paper by Chromik and colleagues[17] who discuss dark patterns of explainability. They present several examples of explanations that do not benefit users but instead deceive or distract people, for example, by overloading them

[17] Michael Chromik, Malin Eiband, Sarah Theres Völkel, Daniel Buschek: Dark Patterns of Explainability, Transparency, and User Control for Intelligent Systems. Workshop on Explainable Smart Systems – ExSS. Organized in conjunction with the 24th ACM Conference on Intelligent User Interfaces (ACM), 2019.

with technical details. When enhancing virtual agents with an explanatory component, we should exploit their potential to communicate information using multiple modalities. Virtual agents also enable us to provide information implicitly. For example, in explainable AI, so-called saliency maps are employed to highlight parts of the input in focus of a neural network. Instead, virtual agents may gaze somewhere so that the users immediately understand where the agents are directing their attention. So far, research in explainable AI that focuses on the end-user is rare. Virtual agents offer the potential to provide explanations in a socially interactive multimodal dialogue.

Stefan Kopp: Another aspect is that to some extent we even capitalize on the non-transparency of our agents. If people would really know how Amazon Alexa works, they would still use it, but it wouldn't be that much fun for too many of them.

Elisabeth André: Yes, indeed. Explanations per se do not increase user trust. The key is the calibration of trust. In one of our experiments[18], explanations helped users choose the smarter agent out of two. They lost confidence in the other agent because they found out how stupid it actually was.

All: Laugh

Catherine Pelachaud: At the same time, there is also this question of transparency, for example, of the mental models the agent has of the users. When the agent interacts the users, it builds a mental model of the users. Do you make this information available to the users?

Elisabeth André: The theory of mind, that's a good point in that context.

Catherine Pelachaud: That's not so easy; it is what the model computes.

Elisabeth André: But it might be interesting to verbalize that: "I believe you would like to do this and that, because if I was in your position..."

Stefan Kopp: In our domains like conversation and interaction, it's often needed to do it in order to resolve misunderstandings, for instance. It is a repair mechanism to use meta-communication which refers to beliefs about beliefs.

Catherine Pelachaud: We should understand if that should be done during or after a conversation. It wouldn't be the same mechanism.

Stefan Kopp: If you use state of the art language technology, that's not how they approach dialog processing anymore, not in terms of what the user or the interlocutor thinks, believes or wants. They have a record of high dimensional embedding of dialogue state. That's not something you could explain easily.

[18] Tobias Huber, Katharina Weitz, Elisabeth André, Ofra Amir: Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. Artif. Intell. 301: 103571 (2021)

Elisabeth André: The benefit of the typical explanatory AI approaches for the end-user is not always clear. Many approaches highlight essential parts of the input to explain how a neural network came to a particular decision, such as nasty words in hate speech detection. Such information might be valuable for researchers who have to tune the neural network. However, it does not suffice to highlight parts of the input on which virtual agents focus to explain the complexity of their behaviors. It's as if I don't understand Stefan's paper, and I ask him for help, and he sends me the document back, where he simply marks some words and says: "Okay, that's my explanation of my fantastic approach."

All: Laugh

### A.3.4 Question 3:

Birgit Lugrin: So about evaluation. How can we actually validate the computation and architectures in the future?

Joost Broekens: It depends. You mean the architectures or the agents that embody these architectures? I think it very much depends on what you're looking for. For some of these architectures, or at least their instantiations, you might want to investigate a particular interaction phenomenon, for example; another may want to study the desired effect in particular use case scenarios. I'm not sure if that's what you mean by validating.

Elisabeth André: We may look at evaluation metrics developed in software engineering as a first step. Here, general quality attributes, such as reusability or maintainability, have been defined. However, we also need to identify quality attributes of the architecture that are particularly relevant to embodied agents, such as how well the architecture supports the implementation of dynamic and reactive behaviors of an embodied agent.

Joost Broekens: But it's a different interpretation of validation. There are many ways for validation.

Stefan Kopp: I think it's both. It's fair to say ultimately, we are going to evaluate agents in terms of the behavior that the architecture is able to produce and whether that actually fits, as Joost was saying, the demands of the application context, whether that behavior is understandable, acceptable, convincing and you name it. At the same time, you can also validate the architecture in more technical terms like how efficient is it, how does it scale up, how does it allow generalization from one domain to the other, can we apply it to both virtual agent as well as to robots, or parts of it. We don't have very good measures for the latter, at least.

Joost Broekens: Many of the architectures are quite modular in the sense that they, for example, have an appraisal module that's connected to a particular BDI[19] engine. But there's also

[19] Belief Desire Intention

a personality module, and others that get all glued together. In one particular scenario, a module does or does not do the job. It's quite hard to figure out what the validity for one of those modules really is, and if it is even relevant. I find it harder and harder to decide on this. This is a really difficult question to be honest. You would want these modules themselves to be software tested, interoperability tested, but also behavior tested. If you enter information into your appraisal module and happiness comes out, you want to be able to check whether that's correct or not in many different cases or whether this was just a lucky case. That's difficult because the modules are not that isolated in the way they're used; they are usually developed more isolated than the real thing.

Stefan Kopp: I think this difficulty to answer this question might be one of the reasons why it's so hard to get papers accepted that propose an architecture.

Elisabeth André: Such papers are accepted when people release an implementation of their architecture to the public. And the impact will be high if the community exploits it successfully. But it usually takes time for the developer's work to catch on.

Birgit Lugrin: It's very important though.

Joost Broekens, Stefan Kopp: Exactly.

Elisabeth André: Since embodied agents are supposed to simulate human-like behaviors, an architecture should be grounded in social and cognitive sciences theories. The implementation of the architecture should not only work. There should also be a social and psychological underpinning.

Joost Broekens: On the output side, I always like variations of scenario-based verification of the behavior. If you test out a couple of scenarios with your architecture, you show what works and what doesn't. At least you get an idea of whether the complexity of architecture is really needed.

Elisabeth André: An architectural framework would be helpful if it enabled us to explore questions we might have on the nature of human-like behaviors. Even if an agent behaves strangely, it might still provide valuable insights for social and cognitive sciences researchers. I would like to have an architecture that enables these kinds of experiments.

Stefan Kopp: Exactly. I was going to say the same thing. Usually, we take the engineering approach. We want to have a good architecture that produces good behavior that meets the needs of the project and the users. But we also have to address a cognitive modeling challenge. We would like to have an architecture that is plausible, like a good model of a theory or understanding, like a cognitive or social science theory. It's also a way to validate the architecture, whether it's a good model and whether it's theory-adequate.

Catherine Pelachaud: Do you remember when we worked on SAIBA at the first meeting in Iceland. Stefan and Elisabeth, you were there. We defined four or five scenarios. We

spent quite a lot of time on defining scenarios that could serve as testbeds to find that if we needed a feature but could also be used for testing. It's quite interesting that would come up with these ideas.

Joost Broekens: I don't know of a set of accepted scenarios for testing novel emotion models. I've been thinking about this for some time. I know that Jon Gratch[20] uses negotiation as a domain to test everything in. You can call it a scenario. But it's not like a validation scenario. That is, if you do this and this and this, this should probably come out of your agent with these probabilities. It's not like you can run your agent against some test scenarios. That would be nice; like a standardized interface you could use.

### A.3.5 Question 4:

Birgit Lugrin: Let's go to the last question. Are robots the new IVAs and how do you foresee the potential of augmented reality?

Joost Broekens: No, and I'm not an expert on the latter. Laugh. I assume we're talking about humanoid robots, so socially interactive robots. I think they won't be a new IVA at all. That's my personal view. We can disagree. We are still very much in search for the killer use cases for social robots in the first place. We've been focusing a lot in social robotics on perception studies with robots and showing that robots are motivating for children, engaging, etc. But I think that if you put them in a usage context, then it's very hard to make a very concrete case. It's actually much easier for IVAs to make many different use cases. They're much more confined in terms of the input / output and of the control you have over them. Social robots, especially social robots that have a motivating and engaging effect for young children for example, that I know most of are lacking quite a lot in terms of social awareness. This is really a difficult issue. IVA might also lack a lot, but there are many use cases you could think of for IVAs that won't hold for social robots that are physically present. The other question, I don't know.

Elisabeth André: I want to comment on the potential of virtual characters for augmented reality environments. Virtual characters that inhabit and thus augment our real environment hold great promise for various applications. Tourism is just one example. For example, users could choose between Cesar Maximilian or the Fugger merchant as a virtual city guide when visiting the old city of Augsburg.

Stefan Kopp: I would also agree with what Joost said in the beginning about social robots. People in robotics used to say that robots create a stronger presence, and therefore they are more effective in terms of their impact on the users. That's true to some extent. They create a physical presence. But the question is what kind of social presence does an agent create.

[20] Jonathan Gratch, University Southern California,https://scholar.google.com/citations?user=HF448PMAAAAJhl, co-author of chapter 12 on "Rapport Between Humans and Socially Interactive Agents" [Gratch and Lucas 2021] of volume 1 of this handbook [Lugrin et al. 2021].

So what Joost was saying is totally true. Robots have yet to deliver on a lot of things that are needed, for example expressivity, fluent interaction, responsiveness. There are still many technical challenges with this. Right now, it looks like that they are not going to be the next IVAs. One could even say that the field is facing a risk of drying out. There's this no real killer application or fully convincing use case of a social robot. My impression is that people are moving from purely social robots to manufacturing or service robots, to try to make them more socially intelligent. This is really promising and important because such robots are starting to enter collaborative settings or assistive scenarios. I think these different branches of robotics are going to merge. Social robotics will inform the development of more socially interactive classical robots for different everyday settings in different contexts.

Joost Broekens: That's a possibility indeed. There's also something else going on with what you said. Many of the social robotics platforms unfortunately are very much vendor locked, even the open ones. But the point is that there's not really a very good content strategy for social robots. I like to approach social robots more as a new interactive medium rather than an agent nowadays. We've focused maybe too much on the fact that this is like an embodied agent, that is supposed to have all kinds of human values. While in essence what we're supposed to be investigating – I mean not us as we're interested in something else – but there should be people investigating how to develop interactive content for this novel medium, just as if it's like a laptop or a TV or a tablet. It has been lacking for quite a long time and as a result, there is no content frame for social robots. And as long as there isn't a lot of content to get on these devices, it's hard to see what can you do with them. Even though you can do quite a lot with them just because they are actually physical humanoids.

Stefan Kopp: But even if they're not very humanoid or anthropomorphic, what's important about them is that they are forced to have minimal social signals, like minimal elements of social interaction. And that's really super interesting. On the one hand, with IVAs you go for the whole enchilada, the full picture you would like to have like a very human-like appearance and expressiveness. This is also interesting but a very complicated and daunting task. I'm quite excited about social VR, augmented reality, or extended reality technology that we have now. I think they can really be a big driving force for a lot of IVA approaches and technologies. There's right now a lot of exciting work on machine learning for social behavior processing, animation or graphics for the purpose of enabling social VR. It's going to be a big playground for a lot of work that we have been doing, also in the IVA community.

Elisabeth André: One advantage of virtual agents is that you can take them easily with you. You don't have to carry a heavy device, but instead, they accompany you on your mobile

device. Virtual agents may also serve as invisible friends who encourage you whenever you are confronted with a challenging situation.

Joost Broekens: I honestly meant I didn't know that much about augmented reality. But I do hope that if there is so much potential that the same mistake will not be made as with social robots; namely that you get again caught in that vendor lock-in. One of the biggest issues that people are trying to now commercially build is pillars. But, it would be great if you could develop a cultural intelligent virtual agent that can give you information and tour guides in tennis, for example. But I would like to be able to run it and download it on any devices. Otherwise I will never buy one of those devices if I am stuck to the software. This is really an issue at the moment. You see that with social robots as well as you saw it with mobile phones well. It is that only when you get standardized platforms that are big enough to support a large community of users, and so friendly fellow person that you will actually get sufficient content for those platforms. Android was like that. It was a silver bullet away.

Stefan Kopp: But then again there are different motivations. Will the technology make it to the mass market. This is one question. The other one is, will it be good technology in order to make scientific progress.

Joost Broekens: Absolutely, absolutely.

Stefan Kopp: In VR, there're big technological advancements. You can have it without markers now. You can have it on a mobile, even with very detailed face tracking. There're a lot of things that we can really envision nowadays there, and there could be great new developments also in our field.

Joost Broekens: I agree.

Birgit Lugrin: Anything you want to add or anything you want to state?

Elisabeth, Joost, Stefen: Thank you for the opportunity.

## A.4 Interview 3: Evaluation

For our third interview, taking place in October 2021, we have drafted the following five questions:

Question 1: Do we need new methods for evaluation other than perception studies? What could they look like? Should we enforce in-situ studies?

Question 2: How shall we define benchmarks that capture the quality of social interaction, or to measure learning gain?

Question 3: How will we conduct and control long-term evaluation and integrate social functions to facilitate adaptation, rapport, or engagement?

Question 4: Shall we share the user models and data used for adaptation with the user to ensure transparency?

Question 5: Are robots the new IVAs? How do you foresee the potential of augmented reality?

### A.4.1 Participants

**Agnieszka Wykowska**, Principal Investigator at the Italian Institute of Technology leading the unit "Social Cognition in Human-Robot Interaction", and adjunct professor of Engineering Psychology at the Luleå University of Technology

**Timothy Bickmore**, Professor in the Khoury College of Computer Sciences at Northeastern University

### A.4.2 Question 1:

Catherine Pelachaud: Do we need new methods for evaluation other than perception studies? What could they look like and should we enforce in situ studies?

Tim Bickmore: Well, I have strong opinions on this.

Catherine Pelachaud: Please go ahead.

Tim Bickmore: I have one foot in the social agents and HRI world and another foot in the medical world. And in medicine, they care a great deal about evaluation, and they don't take anyone seriously until they've done significant large scale, ideally longitudinal, properly-powered, randomized, clinical trials. Most of my funding comes from the U.S. National Institutes of Health, and in order to get funding, you spend half of your proposal writing about your evaluation plan. So it's a very big component of doing research in the healthcare world. These need to be actual in-situ studies; you're putting artifacts out in the world. You have to have some kind of health outcome, on ideally objective, non-subjective self-report measures, some kind of a blood draw, or accelerometry, or some other kind of objective measurement of outcomes. So that's sort of the standard in that world. To me, research is all smoke and mirrors until we get to that point.

Catherine Pelachaud: And you, Agnieszka?

Agnieszka Wykowska: I'm an experimental psychologist. So I would say that for us, it's very important to have well controlled experimental studies. And it is very important to understand that you have to develop mechanistic explanations, which are not necessarily possible when you do things in the field. I'm not sure what you meant by methods other than perception studies? Whether perception like strictly speaking perception, so people being seated in front of a robot and perceiving the robots, or being just in an observational mode, or whether you meant perception as general lab studies? So I would say that we definitely should go away from just perception studies, and we should have way more of interaction involved. Especially when we talk about social cognition, because we know from a second-person neuroscience perspective that social cognition doesn't work in just observational mode. Social cognition is for interaction. So if perception is meant as opposed to interaction, then I would say we definitely need to try to have experimental protocols that are more interactive. If perception is meant as laboratory studies versus in the wild studies, here I think we are still in a phase where we don't understand mechanisms of social cognition, and we need to develop theories, mechanistic theories, to understand and be able to predict what happens. And only then I would say that we would be ready, from science perspective, to bring robots into the wild. So if we want to have scientific explanations, we still need to have well controlled experimental designs.

Birgit Lugrin: Yes, it was really meant that way, perception study versus interaction study, and as a second part of the question whether we should go into the wild more. What do you say about that, Tim? Because you go in the wild a lot, to the clinics.

Tim Bickmore: I always think of studies in terms of the causal chain from the thing that you're manipulating to the end effects, which may be the final health outcomes after a year of interaction. So, at the very beginning of the causal chain is "do people perceive what we want them to perceive regarding the social artifact", "do they perceive our manipulations"?, so simple manipulation checks. "Do they actually interact with the artifact?", "Do they interact with it if it's a voluntary system?" And then, what's the next step in the causal chain? So how does that impact their attitudes toward their health behavior? And then from attitudes, how does that impact their intention to change their behavior? And then from intentions to change, how does it change their actual observed behavior? And after some period of time, how does their actual behavior affect health outcomes, that you're measuring maybe months later. And things can break anywhere along this chain, of course. So perception is important, but it's only the first step, in my opinion. I remember a discussion I had with Stacy[21] about this, that whether a

[21] Stacy Marsella, University of Glasgow, https://scholar.google.com/citations?hl=deuser=LkoaA0gAAAAJ, co-author of chapter 7 on "Gesture Generation" [Saund and Marsella 2021] of volume 1 of this handbook [Lugrin et al. 2021].

character does two or three hand gestures in a given interaction might have very little to do with somebody improving their glycemic control, if they're diabetic, a year after that interaction with the character, right? So some of these things, like focusing on the intricacies of hand gesture for a virtual diabetes coach agent, are just not very important, as far as I'm concerned. Even though I find them fascinating, and am personally very interested in them. But if I'm looking at this from a long-term outcome perspective, some of these things are not very impactful, I would say.

Catherine Pelachaud: But, and I'm going to be on Stacy's side on this, it's true that over a long period of time whether one or three gestures, well I agree it may not have a strong impact. However, the behavior of an agent can be interpreted as a different attitude, or is part of building relationships with others. So, it may have an impact on the interaction. It is this building up of behaviors perception that may have an importance. It is not so easy to control, in-situ and in long-term studies.

Tim Bickmore: I agree that these things can be important. And also very important is getting initial acceptance. So for all of these longitudinal studies, we see declining use over time, for voluntary use systems, for the most part. And so some of these subtleties can impact continued use, which then impact long term outcomes. So they can be important. It's just that the impact of agent nonverbal behavior during one particular utterance during one particular interaction, like the number of gestures the agent uses, may not have much of an impact long term.

Catherine Pelachaud: So how could we merge these up? So Agnieszka's approach is one type of approach because she's really trying to understand the mechanisms, while Tim is working on having an impact having an concrete application in mind. So how could we work on bridging the gap? I mean, how could we use some mechanisms that you can understand, for example, to adapt the agent behavior for your given application? How could we do that?

Agnieszka Wykowska: Well, I think that once there are mechanistic explanations and theories that are based on understanding the mechanisms, then the applications will follow. So zooming out a little bit, first fundamental science and then applied science, right? So of course the problem here is about generalisability of certain mechanisms. So we often have knowledge about the very basic ways that things function and then we need to think about all possible application scenarios. And that's I think where the difficulty comes. So even if we understand how things work, especially when it comes to the human brain, when things work in the lab, that doesn't necessarily translate to real life scenarios. Often in the lab we're lacking ecological validity. That's why I said earlier that I believe that interaction is important, even if it's in the lab. So at least we do have some ecological validity interaction, and at the same time keeping experimental control. So it's a long

process, I think, to get from fundamental research to applied research. But ideally, if we really understood how the mechanisms work, we would be able to then translate them to applications.

Tim Bickmore: I agree. Of course, they're both important. You want to build your agent applications on a sound foundation of results from experimentation and theory building. On the on the flip side though, even within health behavior change, there're lots of theories that are out there. And even when someone says that they're designing an application following some theory, there's this huge chasm of interpretation of how exactly they implement the theory in in their application. There is always a lot of interpretation and subjectivity in the application design process. So it's always very difficult to say whether somebody has done a high fidelity instantiation of a theory in the particular application that they're building. It may or may not reflect the results from prior, more controlled experiments used in theory derivation. If that makes any sense?

Birgit Lugrin: Yes absolutely. I also think both approaches are important if we want to push the boundaries in socially interactive agent research: gain fundamental knowledge in controlled settings in the lab, and understand what actually works out there in the world in a given context.

### A.4.3 Question 2:

Birgit Lugrin: In case you don't want to add more, I would move on to the next question. So we were also very interested in how we should define the benchmarks to capture the quality of social interaction or, for example, measure learning gains or things alike.

Agnieszka Wykowska: I think it very much depends on what is it that you're interested in. So even when you're saying quality of social interaction, what does that mean? Whether you're interested in whether the interaction is truly social, then quality is how social it is, or whether it's a quality of social interaction in a sense of whether it's comfortable, it's easy, it's intuitive, right? So I think that would very much depend on what is meant by quality of social interaction. So if it's about social, whether you want to understand that the interaction is actually social, I guess the important comparison is always a comparison with another human, because that is the truly social interaction we have. I'm always a supporter of objective measures, not subjective questionnaires. So I always try to design experiments in such a way that eventually we manage to have some markers of indicators, objective indicators of interaction. So if I was to think about a benchmark for assessing whether an interaction is social enough, it would probably be a way to measure whether similar social areas of the brain are activated when interacting with a robot compared to a human. That would be like a measure of how social interaction is. When it comes to, let's say comfort, one can look at physiological measures, the degree of stress, or how easy it is to solve another task in terms of cognitive load, and things

like that. So I would definitely be for objective measures. But then the way one would benchmark things would be dependent on what is it exactly that we want to measure?

Tim Bickmore: I agree. In the medical world, there are some measures that assess qualities of doctor-patient and nurse-patient interactions. For example, there's the RIAS, which is the Roter Interaction Analysis System, but that measures things like how many utterances each party makes, how much dominance there is in the conversation, how many questions are asked and answered, and what kind of empathic opportunities are presented and followed up on. So there are some objective measures in the medical world to try and capture some aspects of this. And then, of course, there's relational outcome measures, which are things like trust and working alliance. So at the end of an interaction, or at the end of a series of interactions, you know how well the social interaction leads to some sense of relationship. So I think those are also important. Also, any perceivable social quality of an interaction can be observed and measured, as long as you can get multiple judges to establish inter-rater reliability. But to me again, these are only principally important in terms of how well they lead to some outcomes of interest. There are outcomes proximal to a given interaction that are important, for example, do people want to continue interacting with this agent? Do they come back and again to continue the interaction? Do they comply or do they adhere to the recommendations that the agent is making of them? Those are some of the important outcomes of the social dimensions of the interaction that can lead to longer term task outcomes.

Catherine Pelachaud: So does that mean that for you, the dimension you're looking at and its different measures depend also on the application. For example, if it was more toward education, it would be other aspects you'd look at. So in the long run, it would be learning, self-regulation, this type of thing.

Tim Bickmore: Absolutely, yes.

Agnieszka Wykowska: Yes. I agree with that, too. That it very much depends on the context and application. I was now just having this thought about, again, quality of social interaction. We have studies that show that social is not necessarily always beneficial for human performance. So it can be very well the case that, and in fact our own data showed it too, that if there is a robot that displays a lot of social signals, it's very distracting for a human. So if a task requires focus and efficiency of performing the task, then actually social might not be the best way to go. So that's another example where context matters and actually social might not be the way one would want to go. Just to add on the diversity of contexts.

Tim Bickmore: Absolutely. We did a study that appeared in the last CHI conference[22], in which people could choose to interact with an agent or using a graphical user interface for

[22] https://dl.acm.org/conference/chi

different kinds of health-related tasks. We looked at when they chose one over the other. And basically, if the task was a brief transactional interaction, they did not want to talk to the agent. Personal preference also played a role. Some people just liked talking to the agent. And so then they would, in general bias toward wanting to talk to the agent, if it was a more narrative task. But again, if it was just transactional where they were reporting something, then they didn't want to deal with the socialities, because they're inefficient, right? They take longer to engage in.

Birgit Lugrin: I am wondering, since you both agree strongly that benchmarks should be very dependent on the application and the context of the application, that on the other hand means that it will stay extremely difficult to compare across systems. If everybody is defining their own benchmarks, how will we deal with that?

Tim Bickmore: Test beds. Competitions where you have standard data sets that you're working against. But I can't imagine what that would be like in social interaction with agents, but we could invent something. But I think it's important to fix the context and the application domain to have these conversations.

### A.4.4 Question 3:

Catherine Pelachaud: That's for sure. So the third question is how we will conduct and control long term evaluation and integrate social functions to facilitate adaptation, rapport or engagement.

Agnieszka Wykowska: I would say that's a tough one. So at least from my perspective, saying that most of our work is done in the lab, and in a very highly controlled environment, this is definitely something that is very far away from our approach. It needs to be done, certainly. But longitudinal studies, how does one do that, making sure that everything is properly controlled? As I mentioned at the very beginning, things get very different all the time. So that's a really hard task to understand how things will develop over the time, a longer period of time, and at home in a natural environment.

Tim Bickmore: Well, this is a lot of what I do. So I would say one key requirement that a lot of researchers have difficulty with is reliability. Your agents have to work, and they have to work in the wild, and they have to work for a long period of time. So that requires, unfortunately, that you often have to back off on some of the technology. You have to build simpler systems that you can thoroughly validate, and make sure they're actually going to work for a long period of time. But then that does open up interesting possibilities of actually studying some of these long term issues. How can you change things over time, by having your agents modify their behavior over a series of interactions. So it's an interesting area of research in and out itself.

Catherine Pelachaud: Here we wanted to discuss about long term evaluation, but also, what you had discussed before, when to add social interaction in order to facilitate adaptation, rapport and engagement in a study. So should it be to continue using the application, or to ensure maximum behavior change? The question is related to integrating some social function. As you mentioned before, sometimes it's good to have, sometimes it's not good to have. Social function could be used to enhance engagement, or rapport building, but could be against performance. So in which term can we measure a long-term interaction? In terms of performance , in terms of engagement, rapport, etc?

Tim Bickmore: Well, there're other measures of engagement, for example, the length of any given interaction. If someone is allowed to have an unbounded conversation or interaction with an agent in a given session, you can look at how long those interactions last as a more immediate measure of the quality of the social interaction. In addition to how long they continue using the system over time. You can also look at uptake rates. For example, the agent can present people with opportunities for doing social chat, or off task talk, and see how many of those bids for new topics that users engage in as another measure of social interaction. We can also look at the quality of the user utterances, for example, are they more task-oriented in nature? Or, do they demonstrate more of a social interaction? For example, are they using idiomatic terms of engagement, terms of closeness when addressing the agent, or using close or distant forms of farewell when they disengage from the interaction? Those are qualities of language that you can also look at, to assess how well the social interaction is going.

### A.4.5 Question 4:

Birgit Lugrin: So last question is whether we shall share the user models and the data used for adaptation with the user to ensure transparency? Or shouldn't we?

Tim Bickmore: Well, probably we should, especially if this is something that is put out in the real world. It gives people a warm, fuzzy feeling to know that they could look there if they wanted to. However, it is likely that the vast majority of people will never bother. There's a few people who really care about these things. But to most people, it is just too much work, and they don't have time to bother. But, that's just my opinion. So maybe you just have a button that says "look at the user model", but it doesn't do anything, that would probably work just fine.

All: Laughing

Catherine Pelachaud: That is somehow related to a question we have on ethics. Since the system is building a representation of the users, and the users may, most probably, not know what this representation is. But those representations are being used by the application. So, you may want to have access to those information.

Birgit Lugrin: True, but it might also break the impression of having a truly socially interactive agent when you share the user model, and you see that it is actually rather simple.

Tim Bickmore: But, you know, users don't understand, for the most part, how those models are used. It can be very complex systems that are interpreting them in different ways. So I think it's more about just giving them a feeling of trust, or confidence in the system, rather than actual understanding.

Catherine Pelachaud: Yes, and how can you build the system of trust into the entire system? Not only in the agent, but the entire system? Because as part of a long term study, if you don't get trust, people just stop interacting.

Agnieszka Wykowska: I completely agree with Tim on this point. I think it's probably a feature that is nice to have for the feeling of comfort or trust, but probably doesn't do much in practical terms. And regarding trust, I think it's one of those very big words that are being used often, and they're not so well defined to really understand how to measure trust and how to improve trust. Because what is trust in the end? It has so many dimensions. Is it about trust in terms of trusting the reliability of the system? Or social trust? And probably we trust other humans in very different ways. So with persons, we would trust a person in some dimension, but not necessarily in another, right? So I think it's one of those concepts that is being used in HRI, or HAI, a lot without very clear definitions. And I think it would be very useful to have clearly defined concepts, such as trust.

Tim Bickmore: I agree. It's one of those, what Minsky calls "a suitcase word" that has lots of lots of different meanings, right? There are some really clear ones: I've had a patient who installed a system in her home and came back a month later, and she hadn't turned it on because she was afraid it was beaming data back to the hospital about what she was doing. So there's a clear example of not trusting the system. But there are much subtler versions of that.

Agnieszka Wykowska: But maybe she will trust in how reliable the system beams the information.

All: Laughing

### A.4.6 after the questions:

Catherine Pelachaud: So do you want to add anything? These were the questions we had in mind.

Tim Bickmore: Well, this is about evaluation, so we could talk about attempts to develop standard measures for agent interactions for the virtual agents community.

Catherine Pelachaud: Siska Fitrianie[23] .

[23] Siska Fitrianie, Delft University of Technology, https://scholar.google.com/citations?hl=deuser=ULPLSMQAAAAJ

Tim Bickmore: That's right. So that kind of work is important to do. However, there is a vast amount of work by psychometricians, working in psychology and medicine and educational psychology, who have spent years developing validated measures for many of the things we care about. They do validation studies of self-report measures that involve hundreds, if not thousands of people for long periods of time - to ignore that work is, I think, to our detriment. I still see researchers coming up with their own measures for their own studies. One principle I often see violated, is that you should never invent a new measure in a study, where you're also using that measure as your primary outcome. You should do a separate validation study for it first. I tell my students to first spend a lot of time looking for an existing validated measure that taps into what you want, before you even think about inventing something new, because it's not our area of expertise, and we don't have the resources to properly validate it. So I think there's a lot we can do in practice to improve the quality of our evaluations, by looking more toward what other people have already done.

Catherine Pelachaud: In human-human studies you mean?

Tim Bickmore: Yes, I think just because it's a social agent doesn't mean you have to invent a new measure.

Catherine Pelachaud: Yes, true. Any other recommendation?

Agnieszka Wykowska: Well, from my side it will be something that I already mentioned earlier: we have to define concepts in a more clear way, that are more easy to be operationalized. There are many concepts in HRI and HAI, as we said trust, empathy, engagement, and these are very complex concepts. If I were to be asked how to operationalize those, I wouldn't know how, unless we first come up with clear definitions. So I think the field really needs better, more operational, usable definitions of these concepts.

Catherine Pelachaud: Yes, it's true that if you have a very clear definition, you understand better what you measure.

## A.5 Interview 4: Ethics

For our fourth interview, taking place in November 2021, we have drafted the following six questions:
Question 1: What is an ideal SIA? Should they be the perfect assistant or a companion?
Question 2: Should we as researchers on SIAs go beyond stereotypes, regarding gender and other diversity factors? Or should we follow user preferences (maybe including stereotypes)?
Question 3: How should we draw the line between persuasion and manipulation and transparency, for example, in health related applications?
Question 4: Should SIAs be better than humans? What does it involve anyway?
Question 5: How do we manage dependency and addiction that potentially occur in a relationship with a SIA?
Question 6: How shall we deal with the popular fear of robots overtaking the world?

### A.5.1 Participants

**Ruth Aylett**, Professor of Computer Science at Heriot-Watt University in Edinburgh
**Kerstin Dautenhahn**, Professor and Canada 150 Research Chair in Intelligent Robotics at the University of Waterloo in Ontario, Canada

### A.5.2 Question 1:

Catherine Pelachaud: So the first question is what is an ideal socially interactive agent? Should it be the perfect assistant or perfect companion or something else?

Ruth Aylett: I don't think there is an ideal one. I think the answer to that question is bound to be *it depends*. What are you trying to do? Generalities are a bad idea in AI and robotics. General systems don't work. Specific systems tailored to particular niches, can work and can do useful things, but then you must tailor the SIA to your niche. Which means there isn't an ideal SIA at all.

Kerstin Dautenhahn: Yeah, I would agree it depends very much on the tasks, on the application areas, and also on the people. Is it about children? Is it about adults? Is it about people from specific groups? I think that this really depends. Also, I don't see a contradiction between assistant and companion. Companions can be assistive, and assistants can have an element of a companion, in terms of being a friendly presence. So as well said I would agree. It depends.

Ruth Aylett: So an assistant is a functional term. It describes a set of functional capabilities. A companion is not, in that sense, quite the same sort of term, because it suggests an affective relationship, and something that's involved in your social life, and not just in particular tasks. It's not necessarily a task related term. So I think those are two very different terms. If an SIA has no useful functionality, then I don't see that it will be

accepted in whatever niche you're trying to put it. Because as we've seen from the failure of so many companies, the purely social aspects are poor compared to humans, and they are going to continue to be poor compared to humans for a long time. So if there's no functional aspect to an SIA, I don't see that anyone is going to give it floor space in their lives, whether professional or personal. So I think there has to be some kind of functional capability, whether you call that assistant or something else. As for companion, we're getting into the ethical issues further down about affective relationships and people becoming dependent. So let's park that one for a minute, unless we mean something completely different by companion than we would in the human case. Which is quite probable, I would have said.

Kerstin Dautenhahn: Well, I still think assistants, in order to be successful, in order to carry out their tasks, they need to have some social abilities. Whether we go down the line of encouraging relationships, that's a different topic. But if you talk about therapy, for example therapy for children, you cannot have a robot that is purely functional and just tells children "you have to do this", "you have to do that". No, for that particular application, you would need social engagement. So this is what I meant when I said that sometimes these terms can be overlapping. But I do agree, bottom line is, we need these robots or agents to be useful, to actually do something.

Ruth Aylett: I definitely agree with what you just said. My point was almost the opposite of yours, which is you can't just have social capability.

Kerstin Dautenhahn: Oh, OK. Yeah, certainly I agree.

Ruth Aylett: I agree with you that just having functional capabilities doesn't work either. Particularly in the sorts of niches we're looking at. So, if we don't use those terms, if we use *functional* and *social* instead of assistant and companion, I feel a bit happier with the discussion.

Catherine Pelachaud: We use those terms, as they are commonly used in the literature. But I agree, it's better to view it as a functional task or social.

### A.5.3 Question 2:

Birgit Lugrin: Thanks. So I think we can directly move on to the next question, if that's fine by you. So we were wondering, should we, as researchers on socially interactive agents go beyond stereotypes regarding gender or other diversity factors? Or should we follow user preferences that might include existing stereotypes?

Kerstin Dautenhahn: These days there are a lot of discussions on stereotypes, and also about norms. Should robots follow certain norms? I was part of a workshop where I brought up the metaphor of the 'echo chamber', which we all know very well. And so the question is, do we want to develop these agents in a way that they just match what people expect?

Or do we want to make them into some interesting artifacts that might challenge some of these stereotypes and expectations? I myself would certainly prefer not to have robots that are presented in a very human-like way, for example android robots, as a very particular person with a very particular background and gender. I would prefer more neutral and more challenging robot designs, and to base designs more on artistic skills, rather than just trying to faithfully emulate human-like shapes and behaviours. So I think it's an open question. If people have certain expectations and these expectations are not met, then they might disengage from the interaction. But on the other hand, I think through the design we can also challenge people and go beyond the stereotypes and norms. Which could make human-robot interaction experience more interesting.

Ruth Aylett: I think there's a different answer for graphical characters and for robots. Very human looking robots are a disaster area normally, because they raise expectations about actual behavior that we can't meet. Even if we wanted to. So trying to produce a very human-like robot is going to produce not stereotypical behavior, but bad behavior, which is going to annoy people. Unless in very, very short interactions, like some of these scripted interviews you see Sophia[24] doing. But anything that's truly interactive, where the niche isn't very, very narrow, like one or two interactions, a robot is going to fail if it produces expectations of humanness. Because it won't behave like that, it will glitch, it will fail, and people will just get irritated about it. So unless you are in a deceptive situation, which I would say these Sophia interviews are, I don't think that in the robot case, there's any point whatsoever in trying to make it look really human. And I agree exactly with what Kerstin said: We should look at art, and in particular at animals, cartoon characters, stereotypical machines maybe - depending on people's preferences, and not at human stereotypes at all. So stereotypes are a loaded word, and it's not clear in context, what exactly you mean by it. If we think of graphical characters, we can produce a very photorealistic face and so on. Although again, its behavior will not necessarily match that appearance, it won't be as fallible as a robot would be in the same situation. The ideal of young, attractive women has passed its sell-by in that particular area. I would like to see a little less of the routine assumption that that is the way graphical characters should look. For example how about women that are not 'beautiful'? If you're going to have something that's gendered, then women that are older, or some male characters. In principle, I think Kerstin is right. We would be better not having to gender the characters, but it's almost impossible to stop people doing it. We try. We have a robot called Alyx[25], which is the Emys[26] robot. It doesn't have any amazing human-likeness and it doesn't have any real gender cues. But people still gender it, in spite of our best efforts. So I

[24] https://www.hansonrobotics.com/sophia/

[25] https://www.socoro.net/

[26] http://doc.flashrobotics.com

don't think you can stop people gendered SIAs. You have to be aware of that. I came up with the idea of social affordances when I was talking about this in my book[27]. So there's a chapter about appearance, which is one of the things we're talking about here and the idea of social affordances, which is: you look at a social agent, what do you think it can do from its appearance? A very human-like appearance gives you one set of social affordances, a machine-like blocky tin person, gives you a different set of affordances. And these are expectations. So you could say they're stereotypes that the user is then going to apply to the interaction. But you've also got behavioural stereotypes, and I worry about those too. So do we make these characters submissive? Do we make them overhelpful? That's where Amazon's Alexa went wrong, I think. It's got the persona, though not the body of an overly helpful young woman. And as a result, in conversation you get sexual abuse, and get some really horrible dialogue. And that's partly because of the way it taps into the misogyny of the people who are interacting with it, because it's an overhelpful young woman. So, yes, we've got to be conscious of these things. What social affordances are we providing and to whom?

Catherine Pelachaud: People have worked on the question of the voice. How to make a genderless, non-gendered voice?

Ruth Aylett: Well, a low woman's voice is not very different from a higher man's voice. Certainly, I've been mistaken for a man many times on telephones. So we can make the voice a sort of alto which could be gendered either way. It's not difficult to do that. It's a little bit different when you're using unit selection voices because these are based on real people, and they tend to accentuate the particularity of the person on whom they're based. So if you look at a male unit selection voice, it does tend to be very male and the female voices do tend to be very female. But that's because they picked people who have rather identifiable voices. And that's a choice. You could get people whose voices were not quite so gendered, if you wanted to.

Kerstin Dautenhahn: If I think back 20 years ago when we tried to make robots speak, there were only very limited choices on the voices. And sometimes we just ended up with a horrible voice because it was just what we had. But these days I usually encourage my students, if they have to pick a voice, try to make it as gender neutral as you can. And ideally also with a little robotic aspect to it, to at least not obviously invite people to treat it as if they are talking to a human being. And similar to what has been mentioned before by Ruth, there is the tendency to anthropomorphise. Regardless of what you present to people in terms of how the agent looks like, or how it behaves, people will end up saying 'he' or 'she' in my experience. I always say to my students, when we write a paper, it is 'it', the robot. It, not he or she. And that's also how we should talk about them. But it doesn't

[27] Aylett, R., & Vargas, P. A. (2021). Living with Robots: What Every Anxious Human Needs to Know. MIT Press.

really work in my research group. I'm not very persuasive, it seems [laugh]. Whenever I have a meeting with my students and they talk about the Fetch robot[28], for example, a very mechanical looking robot, they often say 'he'. And when they talk about Pepper[29], they often refer to it as 'she'. And then I have to remind them it's a robot. In society in general we are not encouraged to think more about gender fluidity, but it's very difficult to get that into our daily, day to day spontaneous language when talking about robots.

Ruth Aylett: Very difficult. You raised a very interesting point there about voice. I think there are specific issues to do with voices. We as researchers have thought about appearance over many years, quite a lot. And we have thought very little about voices. So Pepper has a voice, which in my experiences, if you interact with it over a long term, it begins to get on your nerves very badly. And now, even more so, those voices are not particularly gendered, or they are childlike, if anything, in pitch. But they're not very comfortable to interact with over extended periods. We can use unit selection voices instead, and they are much more comfortable to interact with over long periods, but they do carry human baggage. So we have to ask ourselves, do we then say we shouldn't use unit selection voices, even though these are more comfortable for people over longer term interactions, because they carry baggage with them? Should we go back to horrible synthetic voices which people are not going to want to interact with over long periods, because they're just too irritating after a while? They do work in a demo but there's a difference between what works in a demo, and what works over the long term. An itchy voice works well in a demo. It's kind of, wow, it's a robot, it's got a voice. And then if you do it for three days, it's like, oh God, that robot gets on my nerves with its voice.

Kerstin Dautenhahn: Just as a side remark: my team created a voice for Fetch and they decided, I was not involved in that decision, they wanted Fetch to speak with a British accent. They could choose between American, Canadian or even French English accent. And now they sent me a video of it. I found it really, really interesting. You have this very mechanical looking robot, it's a more industry type robot, very strong, too. It's certainly not what you would expect for a social robot. And then it speaks with this perfect British English. And I found that interesting. But as Ruth said, of course we need to see how in more repeated interactions that might work. But in a short term interaction, it certainly creates this level of surprise. I mean, I was surprised when I saw it. A Fetch robot and a British accent, where does that come from? It was kind of surprising and then questions people, it doesn't look anything like a human and it doesn't look very social, but it speaks in this way. But we need to see, because long term interactions are really key, and it is also in long term interactions that people might realize how useless a robot is.

[28] https://fetchrobotics.com/

[29] https://www.softbankrobotics.com/emea/en/pepper/

Ruth Aylett: I mean, the voice carries more information than gender. So particularly in Britain, the way in which you speak, gives you an indicator of class as well, typically. So never mind just using a British, I would say Received Pronunciation (RP) voice, which may be what you're talking about here, you can also give a robot a regional UK voice. So we've used a Scottish female voice on our robot typically, not a hard-to-understand Scottish, it's what I would call middle-class Scottish, but it's clearly Scottish. We could have given it a Glasgow male voice that would give it a different impression. In either case, the voice being better than the appearance, or more human-like than their appearance makes people think better about it than if it were the other way round. So we've done experiments on voice, and there's something to do with the relationship between the appearance and the voice. We haven't quite got that. We did an experiment with the NAO[30] on the Emys with children, where we swapped the voices round. We ran them with the voice they came with and we also ran with the opposite voice swapped round. Then we asked people to evaluate them, and it turned out that they liked them best the way round that they were originally. They liked the NAO more as a sort of friend because it's more childlike. I think the Emys carried more authority, and its voice was an adult voice, so that also carried authority to children. And they didn't think it was as friendly as a result. And they didn't like it very much with the NAO voice and vice versa. They didn't like the NAO very much with a Scottish female voice, either. So there are social affordances, again. You're creating expectations and you're creating an image in people's minds of what this thing is and what its abilities might be. When you give it all of these things, these bits of behavior, and voices are much more important than I think we've really investigated.

### A.5.4 Question 3:

Catherine Pelachaud: So shall we go to the next question, which will be how should we draw the line between persuasion and manipulation, and transparency, for example, in health related applications?

Ruth Aylett: Well, that's a good one, isn't it? Which is where the ethical issue becomes rather pronounced. Normally you would draw the line, if you were a person, by telling people what you're doing. So you would say, I'm going to try and persuade you that... Rather than being like a sneaky advertiser and trying to get underneath your psychological defenses. Humans do this kind of thing all the time. I agree that we shouldn't try to manipulate people with our artifacts. I don't think that's really very ethical. Even if we do it as people. I don't think we should make a habit of doing it with our artifacts. In which case you have to tell people what the system is, and what its limitations are. They have to know that it's a piece of software and hardware, not a person, and they have to

[30] https://www.softbankrobotics.com/emea/en/nao

know what its aim is in relation to the interaction. And even if that breaks the interaction a bit, I think you have to tell people what you're doing. I don't think you should deceive them about this.

Kerstin Dautenhahn: Yeah, I think there's also a very fine line on what do we mean by manipulation. If I give the robot a certain shape, or behavior, or voice, or facial expression, and therefore make people think the robot cares. For example, if the robot laughs when someone tells a joke, or has an empathic expression and gestures and body language in a healthcare context, after all, this is faking, right? I mean, some people might disagree with me, but I don't think robots have genuine emotions. They can, of course, simulate them. You can use them in order to control their behavior. They can certainly express a lot of emotions. They can perceive emotions from people. We have more and more sensors now, not only vision, physiological sensors, for example. They can know a lot on what people are doing and what states they might be in, and maybe detect their facial expressions, and then they could respond in a social way. But as far as I'm concerned, this is also a level of deception here, because this is modeled according to human-like behavior. But the robots are not human-like. Robots do not actually know what pain is, or what love is, or what disappointment is. They can pretend really well that they do. Of course, I would completely agree that no one should intentionally deceive people, or manipulate people. So in a way, robots should be "better" than humans or more toward the ideal on how we think we should be. We should be honest with people. We should clearly say what our intentions are, and what our goals are. But sometimes you cannot avoid deceiving people, although you don't want to. But for me, deception is already if you encourage people to create a mental model in their mind about what the robot is, what it can do, what it cannot do, that doesn't really match their real capabilities. And that's really difficult. But it's something I've seen in the last few years. There's a lot of literature now in the field of HRI in particular, about transparency and also explainable AI, which is related to that. So we need to see how that develops and whether any guidelines will come up. Researchers have already been developing some guidelines about transparency.

Ruth Aylett: They have, yes. There are people in Britain who did this with the funding organizations. There's a set of principles to do with robotics, which seem very sensible to me[31]. You can overestimate how successful these systems are. So this is an interesting question. In practice, the ability of any of these systems to be sophisticated enough to manipulate anybody is a bit limited. So I would disagree, Kerstin, that our systems recognize emotions. They're not even good at recognizing social signals, never mind what the emotions are behind the social signals. That was our experience when we researched our empathic

[31] Joanna J Bryson, The meaning of the EPSRC principles of robotics, Connection Science 29(2)

robot tutor[32]. It was really quite poor at this, and we were using physiological sensors, and we did have a strong context, which was a teaching game for map-reading and it was still very difficult. So I think you can exaggerate how successful these systems are going to be. The problem you've got is that people want these things to be human, just like they gender them. So if you look at the original video, I think it's called Mechanical Love[33], where a long time ago, a Danish author in a wonderful documentary looks at the Paro robot[34] and looks at the rather arguable things that Ishiguro[35] does. Without much commentary, she sits in the background. She follows the Paro's testing in a German old people's home. They test it with a lady who actually doesn't have dementia, I think, but that she hates the home becomes very clear. She hates the home, and she hates the other people. She's really miserable. They give her the robot. They unzip its belly. They show her it's a robot inside. Thereafter she treats it exactly like a living thing. In spite of the fact that they've done the ethical stuff, they told her, it's not a living thing. And in spite of the fact that it doesn't do very much, it doesn't talk, it hasn't much in the way of social behavior, just a bit of face following and a bit of recognizing the sound of one's voice, and a bit of response when you stroke it. She treats it like a living thing, much to the irritation of the people in the home. It's quite funny, actually. So even if we tell people this stuff, they will still impute this behavior to our systems, like gender. And there's not a lot you can do about that: if you like, people are manipulating themselves, because we all have these strong theories of mind. I mean, come on, we impute intentionality to printers and photocopiers, never mind to something which talks. So it's going to be very hard to stop people doing that. Yes, I agree. You have to tell people what your limitations are and probably at regular intervals over a long period, but over a long period, people are going to notice the limitations, believe me. OK? You might get away with this for half an hour, but you're not going to get away with it for a week in the current state of the art.

Kerstin Dautenhahn: I'm just sometimes getting worried about this aspect when people try to encourage affective interactions, because this is in a way, different. I do agree people also talk to their printer or swear at their laptop if it doesn't work. But they would have no problem buying a new one. If it breaks, they would not shed a tear, other than about the costs of the new printer. They can make the distinction between "Oh yes, I sometimes treat it as if it were human-like, but actually I still know it's just a printer", but I think this is different. Maybe in this care home example that you mentioned, and from what

[32] Obaid, M., Aylett, R., Barendregt, W., Basedow, C., Corrigan, L. J., Hall, L., ... Castellano, G. (2018). Endowing a robotic tutor with empathic qualities: design and pilot evaluation. International Journal of Humanoid Robotics, 15(06), 1850025.

[33] see https://www.imdb.com/title/tt1186021/

[34] http://www.parorobots.com/

[35] Hiroshi Ishiguro: http://www.geminoid.jp/en/index.html

I hear Paro is indeed really successful in many care home settings, but this is different. I think Paro is encouraging people to treat the robot like a living thing, a pet-like robot. So I find that very worrisome sometimes. Although I can clearly see that there could be really good benefits.

Ruth Aylett: It works well for what it was designed, which was interaction with people with dementia who are not good at recognizing individuals anyway. I think the problem with a system over the longer term is that you invest it with a memory of your interactions and with a perception of its pattern of behavior over time, which we would call personality if displayed by a human. And that would seem unique, and therefore you would be concerned about it being killed and vanishing from the world. Of course, there's no reason why we can't download all of that into the next one, and it becomes resuscitated. It becomes a reincarnation of the original one. We probably have to think about things like this. That's not human. And I think people will adapt their perceptions if or when they get to do long term interaction with these things. I think at the moment, they tend to be knocked out by the short term interaction. They don't think about the long term. It would be interesting to study Paro over the long term in care homes, if anyone's doing that, and to see what its impact is, not just on the patient with dementia who has little memory, but with the care home staff. People are running the care home, people with longer memories who will remember its behavior and may attribute personality to it from observation. I think we need to know what people think of these things over the longer term, indeed.

### A.5.5 Question 4:

Catherine Pelachaud: Let's continue with the next question which is: Should SIA be better than humans? What does it involve anyway? This question relates to the notion of the agent being perfect, ideal, etc, but also it relates to the relationship the agent builds with the user. For example, we say the agent should be engaged, it should adapt, show empathy toward the user. The question is to be taken in those terms.

Ruth Aylett: Well, they aren't going to be better than humans. Let's be right about this. But as a goal, that would be a pretty implausible one. They are going to be a lot worse at everything. I think it depends on what you mean by better than, doesn't it? If you want to say, should they be more ethical and should they be nicer to people than humans sometimes are, then probably the answer should be "yes", we should produce good behavior. After all, we're engineering this. If it behaves, as we would say in the human case, badly, then it's because we've engineered it to behave badly. And then there's an ethical question. When we were doing our empathic robot tutor, which we designed to help learners with boredom and frustration, I said "well, the educational application is too easy. So we're not seeing enough boredom and frustration. We need to make it

harder. So learners get frustrated and then we can deal with the frustration". And some of the teachers said: "well, that doesn't sound very ethical". My suggestion had a good reason behind it, but I could see their point. So in principle, depending on our niche, we probably do want the thing to behave more patiently, more ethically within its limits, which are going to be severe, and just not get very angry or very grumpy or any of those affective states without a very good reason which had better be ethically determined as well. However, it will fail. We know that. Of course, it will fail. It won't behave deliberately badly. It just won't behave properly, as intended. People may well interpret this as bad behavior. But it's just the fact that these artifacts are not very good, currently. We know this.

Kerstin Dautenhahn: Now, when we previously talked about stereotypes, and I said that, maybe, robots can be better than people. What I was really referring to was the more ethical behavior, not better in the sense of physical capabilities or cognitive capabilities or even social capabilities. But more in the sense of we, as engineers, certainly don't want to put in stereotypes or biases intentionally. I'm not sure if we can actually avoid that, when we use AI, for example. But better in the sense of making the robot more rational. I'm surprised that I'm using this word. But I think actually, in particular now, the pandemic is really affecting people on more than one level very, very deeply, and I sometimes see behavior that I wouldn't expect in a more normal context. I think it's because of all of these effects and because we are people, we are affected by everything that happens. I've now stayed mostly home for almost two years. For a significant percentage of my children's lifetime, they have mostly stayed home. And so it is having an effect. People are affected by everything that happens to them, the people they talk to, what they're reading, what they experience etc. And so robots, of course, since we programmed them, can be made in a way that is less irrational in the way how sometimes people behave. As Ruth said, I would not want a robot to suddenly shout at a person because they get really frustrated. The robot gets frustrated with the person because the robot may try to convince the person: "Oh, you should do this or you should do that". The person doesn't do that. A human would, maybe at some point if that continues over time, get frustrated and maybe even behave in a way that is not so nice, but robots definitely shouldn't do that. They should be more balanced in how they react to things, and they should always put the welfare of the person that they are supposed to assist or be a companion of first. They should always have the person as their first priority. I'm not so much worried about how the robot feels and more concerned about how do the people feel that the robot is interacting with.

Ruth Aylett: The robot can't feel, I agree with you. It has a model of emotion that's all, and just like a model of rain isn't wet, a model of emotion isn't emotion. Nevertheless, people interpret it as having these internal states. Here's an ethical question we haven't posed

and certainly haven't answered as researchers. Why do we allow these systems to say 'I' when we know they don't have an 'I'? 'I feel', 'I say', 'I think'. We allow our systems to say these things. They don't do any of those things. Those are deceptive statements. We do not have to do that. So, in the back story that Tim Bickmore and colleagues' patient robot[36] told people, which improved interactions incidentally, for post hospital treatment, they gave their character a back story and this made it more successful. It said 'I like' various things and chatted about them. Of course it didn't. It doesn't. It can't. We could get around that by saying 'In my memory' or 'I've been given the ability to' and we wouldn't want to do that, would we? Because we'd say that will break the interaction. So, we're on the knife edge here between good interaction and deception. These systems do behave differently from humans and in some ways better. So the robot Paro is virtuous. It never gets upset with people with dementia at all. It will carry on doing what it does. That makes it of limited use to people without dementia, incidentally, as they will soon get the idea of what it does and lose interest. People with dementia don't lose interest because of their memory problems. They still find it new and interesting. It's like ex-US President Ronald Reagan said 'every six seconds, old people I know are new people'. Same thing here. It behaves better and certainly behaves better than a real animal. It will never bite people. It's never sick on the carpet - like my cats all too often are. It's better than a pet would be. But it also behaves worse than a pet would, because it doesn't have the same variability. It doesn't have the same repertoire of behaviors. It doesn't have the same selfhood. You can't look into its eyes and know that there's a personality in there, because there isn't. So it depends what we mean by better, doesn't it? More patient? I think robots should certainly be more patient than people, even if their model of frustration is such that it's getting very high. The Action Selection System should never then come in and say 'when frustration gets this high, shout'. That's a programming choice which we should never make. So, there are some ways in which it should behave in ways that humans would not, when it comes to negative social behavior.

### A.5.6 Question 5:

Birgit Lugrin: Talking about pets and about the companion robots it actually neatly leads to the next question: How do we manage dependency and addiction that can potentially occur through the relationship with those socially interactive agents?

Kerstin Dautenhahn: I'm not sure about addiction. Do we have any examples of people being addicted?

[36] Bickmore, T., Schulman, D., & Yin, L., Engagement vs. deceit: Virtual humans with human autobiographies. International conference on intelligent virtual agents, 2009

Birgit Lugrin: We have smartphone addiction, for example. People see their smartphones as their companion and they get really addicted by them. That could probably potentially happen with interaction with your robot or with your virtual agent as well.

Catherine Pelachaud: There was this example, you recall, the Tamagotchi. That was not really a SIA.

Kerstin Dautenhahn: I wouldn't call that an addiction. These days, people are also very fond of computer games and so on. But robots are still very, very limited in what they can do. For example, let's consider the new version of the Aibo robot[37], which wasn't produced for several years, but now there's a new generation of Aibo robots. In Japan, there are some places where, if your Aibo robot breaks, you can go there and it will be repaired. And people actually prefer to have their old robot be repaired rather than getting a new one. So whether you call that addiction or just an intense attachment to an object, I'm not sure. Addiction would mean that people then don't do what they would normally do. That's my interpretation. Many children these days, for example, play a lot of computer games. Most are not necessarily addicted to them, although they might try to play as much as they can. But it's probably OK as long as they are still doing all the other things that they should do as a child, doing homework, playing sports and so on. So with dependency, I'm more worried about dependency for vulnerable people. I'm not so much worried about healthy adults. I'm more worried about, for example, people with dementia or children where the boundary between, is it a robot or is it a living thing, might very easily get confused. That's where we need to look in much more detail on how do we want to design the robot? Because dependency clearly is not healthy, let alone that these robots at some point will break. If a person thinks of the robot as a living thing, then they would be devastated if that robot breaks. So that's certainly something we need to avoid. But given the state of the art of social robots in particular, I don't think there is a danger of an addiction imminently coming up. But clearly, it's about the ethics of how people who design those robots and programmed these robots, what type of attachment or human-robot relationship they want to encourage. This might lead to dependency, and it might lead in future to addiction. At the moment, the field has more of the opposite problem, namely that as soon as people, in repeated interactions, interact with these systems, they often lose interest. So it's the opposite of addiction. But addiction to robots is certainly something we should avoid, like with any type of addiction.

Ruth Aylett: I think we have to be careful about over-stressing this. I agree with your practical remark. The chances of this happening are low. Paro might produce dependency. You could always say its behavior is not so sophisticated that you could tell one Paro from another very, very easily. So I suspect that replacing one Paro with another Paro in its

[37] https://us.aibo.com/

intended audience probably wouldn't be very noticeable. When it comes to children, we should be aware that children are already dependent on things, like favorite toys, because they impute personality to them. Many children will be very upset indeed if they lose a favorite toy, particularly when they're very young. You may have to buy them another one and hope that they don't realize you've done that. This happened to one of my grandchildren who lost his favorite soft toy; his parents bought another quietly when they couldn't find it. They also have pets in a lot of homes. Pets die on a regular basis. Children are devastated when their pets die. But we don't stop people having pets because of this. So I think we have to be realistic about some of this. The other thing is, I don't think people are addicted to their phone, they're addicted to some of the things behind the phone. The actual physical piece of kit gets replaced on a regular basis. You stick a cable in there and suck everything down into another phone. And there you are. You've got your phone back again, haven't you? Because it has everything that previous phone had. Hence my remark about resurrecting robots. We could easily do that with robots, too. Certainly graphical characters, they never die. You can just transport everything that was in one into another so they don't have to die at all. You can replace them; in the sense of their long term interaction, any information they've acquired and all the rest of it. Just the hardware would need to be changed. There are ways around this problem. I don't know that we want to encourage dependency necessarily. I do not think you can resist it. People will or will not become dependent. But I don't think we necessarily want to encourage that, but we will have to deal with it. We will have to deal with this issue of what happens when it breaks down. We're not going to stop people again. You can't stop people. So let's resurrect the robots so that when one of them goes, you can suck everything down into another one just like your phone.

### A.5.7 Question 6:

Catherine Pelachaud: Shall we go for the last question: How shall we deal with the popular fear of robots attacking the world? World meaning our jobs or not control politics, police.

Kerstin Dautenhahn: Robots that take over the world and wipe out humanity, that's certainly something you could only laugh at. But what you mentioned robots taking jobs, this is a real fear and needs to be considered really, really seriously. Whenever new technology has been used on an industrial scale, it certainly led to changes, whether fewer jobs or different jobs, because then suddenly people were not doing what they have done for many, many years. They were replaced by machines, not necessarily by robots. First it was just mechanical machines powered by steam engines. They similarly also took over jobs that previously people did manually. Now, there's a lot of discussion in the field: will robots also have a big impact? I'm actually more concerned about AI personally. I think AI takes more jobs than robots actually do. When it comes to actual agents, either virtual

or physical, it's certainly something we need investigate. We need to look at what the system should actually do, what we should promote. Is it about replacing what people are doing or is it about providing tools that people could use? In my own work, for example, I emphasize very much when it comes to, let's say, therapeutic applications that we are not trying to replace a therapist or a teacher or a parent. We are providing tools that in the hands of the people, could be used in a therapeutic context, and I'm always very deliberate about that. Whenever one of my students has an idea: "maybe the robot could do what the therapist would do". I say: "Well, we have to be careful here". But, of course, that's my personal choice. There are already lots of discussions also on robots and automation. We clearly need to keep this conversation alive and also involve and inform the public. I'm more concerned about media stunts with robots, like with Sophia. Someone just sent me a link to this interview that she gave. I looked at it and I thought, this is literally not possible. This must be fake. Then I looked into it. And yes, it was scripted. Of course [laugh] it was scripted. But they didn't tell people. You had to specifically look for some information or talk to people who actually know that robot. I knew when I looked at the interview this cannot be real because I've seen the same thing happening with Pepper. I was on a panel at some point in the UK where Pepper was introducing the moderator of the panel they started making jokes back and forth. But in the background, there was someone from Aldebaran sitting there, desperately typing, typing away so that Pepper could give the appearance of having a really meaningful, complex, real time dialogue. But the coordinator of the panel was very responsible. He disclosed that to the audience. He didn't let the audience leave thinking that Pepper actually managed this super interesting dialogue. No, he said: "Look, actually, there is a young man in the back basically puppeteering the robot". I'm more concerned about people's opinion, the public's opinion on robots which is so much influenced by media stunts. It always makes me quite frustrated, actually. And I don't know why people are doing this. Well I can guess why people are doing it, but I'm not very happy with that.

Birgit Lugrin: Movies are probably contributing to that as well, right?

Ruth Aylett: Yes, movies also contribute. If you ask people what their actual experience of robots has been, movies are one of the big components of that actual experience. There are also the Boston Dynamics videos, which are also largely tele-operated or scripted, and had a big influence too. People should know better including the stunts that Kerstin mentions. So one of the 10 principles that the researchers in the UK came out with, that I mentioned earlier, was to have the moral duty to correct wrong statements in the press. I'm afraid we do have to do this. We don't because we would rather research; we just sit there and fume at the stuff that comes out, which we know to be absolutely wrong. But we do actually have to start responding and saying that it's wrong. I've done this quite sharply

in some cases. for example with AIDA[38]. This is supposed to be an artistic robot. It has an A.I. paint program attached to it with some graphics processing. Anyway, I won't go on about it, but I wrote a piece on Medium about it. I also wrote a very sharp piece on a blog from someone in the poetry field. AIDA had a poetry generation program attached to it, so "she now writes poetry" (ironic voice). Hmm. I was fairly shocked about that. I wrote a book with a colleague specifically to counter this stuff, a popular science book we've just brought out [39]. But I will tell you that that is never going to get mass circulation because everyone prefers the alternative story. It's much more fun. People don't particularly want to have their illusions dispelled. They would prefer to believe the hype: it's interesting. It's fascinating. Maybe a bit scary,the same way that zombies are, that robots will take over the world. There are cultural reasons for this, at least in Western societies. So it's going to be very tough. The reason you get the stunts is because people like them, and because it makes people money. No newspaper ever lost money by printing a story about robots taking over the world. People want to hear these stories. They do not want to hear that robots will not take over the world and are really quite clumsy and not very, very useful at the moment. That's not much fun as a story. Why would we tell people that? That's not going to sell newspapers. This is technology hype. Yet we have an active duty to pursue it rather than just sit there and view it. On the automation thing, I think Kirsten's really made the important point here. This is part of an overall process of automation that started in the 1750s in Western Europe, which produced a severe trauma for the populations that underwent it as their old society was torn to ribbons and they were herded into insanitary cities and chained to factory machines. And that trauma echoes down the centuries since. Not for nothing did people try and break the machines at the time. Robots are just another element in this story as far as people are concerned, and that's why they fear them. Should they fear them? Well, automation is a continuing process. I'm not even sure I agree with Kirsten that AI is more of a problem. There's not an awful lot of AI in most of the systems at the moment, minimal amounts of AI. The problem is often, as someone in Edinburgh told me, that these systems are not very intelligent. In fact they're very stupid. But what you've got is hype, which tells everyone who uses this stuff it is really cool, it's really infallible, you can believe in it and you should use it. That is very dangerous indeed. Much of the stuff about robots, incidentally, is not about robots. I've been following some of these stories. They do not mean robot, when they say robot; as we would understand the term. They mean the internet, actually. They don't distinguish. What you're talking about is information automation, which is the current wave of automation. So in automating aspects of information processing on a geographically extended scale that we couldn't

[38] https://www.ai-darobot.com/

[39] Aylett, R., & Vargas, P. A. (2021). Living with Robots: What Every Anxious Human Needs to Know. MIT Press.

do before we have the internet. But the hype is serious. Not primarily because they're knocking out jobs, but because they're overestimating the capabilities of what they can do, and they're excluding human judgment from processes which should not have human judgment excluded from them. So the computer says "no" written large everywhere, in very sensitive discussions. For example the system in the US for sentencing, which helps judges decide whether a particular individual is likely to be a recidivist, break the law again. The sentencing system is trained on a database, which is completely skewed by the high proportion of black people in the US that have already been convicted of crimes because the justice system is racist, basically. So of course, it's biased when it makes decisions. If it's a black guy, it will tell them that his likelihood of breaking the law again is high because it's using biased data, but "oh, the computer must be right. This is an intelligent system" (ironic voice). So the judge will take its advice. The problem you have here is not necessarily the systems. It's the illusions that people have about these systems, which are very deep indeed. For facial recognition, a UK police force used it in a football stadium of people and they got a 45% error rate in identifying people on their database of criminals]. Well, they were lucky, it wasn't higher than that is all I can say. Because if it's 95% accurate under good lighting conditions and you've got 10,0000 people or 50,0000 people, then you're going to get very high error rates. "But it's a technology that works", they've been told. So I think our problems are there.

Kerstin Dautenhahn: When I said I'm more afraid of AI, I didn't mean AI in robots. I meant the A.I. that is used for example in law enforcement surveillance, the AI that's running on my phone that I might not even know about. But that AI knows a lot about me. Although I try proactively to switch off as many features that I can. But I'm sure I overlook many of of those. This is more what I'm afraid of. This is more what is taking over the world. Surveillance cameras everywhere knowing our every single move; also initiatives to build smart cities. Here in Ontario, before I moved to Waterloo three years ago, they were in the process of approving a smart city within Toronto with basically, 24/7 continuous surveillance. I was very happy, when two years ago, they actually scrapped those plans. There was very strong opposition to collecting everything from within your house, outside, on the street, where you go, what you shop, who you are etc. But in some other countries and other places, people already have that. They have very, very high surveillance. These are more the things that I'm concerned about and not so much whether robots will take over the world.

Ruth Aylett: I agree.

## A.6 Interview 5: Ethics in the application of SIA for children with Autistic Spectrum Disorders

We have organised a specific interview to discuss ethical issues that arise when modeling SIAs interacting with children with Autistic Spectrum Disorders. It happened in December 2022. We discussed the following questions:

Question 1: What are the ethical issues related to technological development, commitments announced to family members (that are not maintained due to the complexity of computational development), and institutional use of socially interactive agents?

Question 2: How should we draw the line between persuasion and manipulation that we sometimes need to get certain effects that are desired, and transparency in ASD related applications?

Question 3: Should SIA be better than humans?

Question 4: What does it involve for SIAs to be better than humans?

### A.6.1 Participant

**Jacqueline Nadel**, emeritus CNRS Research Director at La Salpétrière Hospital, Psychiatry Department

### A.6.2 Question 1:

Catherine Pelachaud: The first question we would like to ask you is: What are the ethical issues related to technological development, commitments announced to family members (that are not maintained due to the complexity of computational development), and institutional use of socially interactive agents?

Jacqueline Nadel: I would like to say first that for an institutional use of socially interactive agents, you should obtain the permission of an ethics committee. The ethics committee will ask questions about physical security (for instance, for robots, they cannot be broken into small pieces that can be swallowed or eaten, temperature should not be too hot or too cold, electric elements should be secured, the SIA should not fall easily, ...). I am thinking of NAO for example. NAO falls easily. Also the psychological aspects related to dependency and addiction should be considered. Questionnaires to families and experts as well as results of pilot experiments can be asked to document these points. The ethics committee will ask for explanations about the objectives and procedures that should be offered to the users or their families if they cannot decide by themselves, so that the decision is taken with full awareness. The family should be informed of the benefits and shortcomings of SIAs. There is no illusion. You should say to the parent that it is not magic. The use of social agents will not change totally the specificity of the person. It is important to immediately notice the regularity of use is very important in a situation,

so as to develop good routines in the use of SIAs. Specially for virtual reality, the ethics committee will ask questions about the feeling of presence which is a very important factor if you can do as if the virtual was real or not. A special problem is now about mixed reality when you have the virtual object in the real room. This is a very big problem for people with autism who have not developed false belief. In this case, it is very difficult for them to tell the difference between real object and virtual object in the room. They are afraid of the situation because they don't know what this object is doing in the room they know. So this is a very important and novel point for the ethics committee. For the family, the most important element for me is for the parents to be free to meet the SIA and start interacting with it or to observe their child interacting with the agent. This is the best. When this is done, usually the parents feel confident with the situation, the objectives, with the way their child will behave. It is a very important element for the ethics committee that the person can see the design and be aware of the different elements of the design. I think that if this is done there is no more ethical issues that are worth to be developed.

Maybe I should speak about the fact that usually the ethics committee asks that experts in the domain of autism and also associations of parents of people with autism have already seen the design of the SIA.

If I look at what we did recently where we built a virtual platform for children with autism with a collaborative agent, we had the ethics committee ask us questions: did experts in the field see the material? Is the material secure? How did you use the material? And also we started with neurotypical children to see if there was a problem that people can explain. This is part of the ethical issues but it is also part of the design of the research itself. So it is not easy to distinguish what is ethical and what is the research itself.

If you would like me to develop more, please tell me.

### A.6.3 Question 2:

Birgit Lugrin: Thank you, that was very informative. Maybe you can elaborate a bit more on how we should draw the line between persuasion and manipulation that we sometimes need to get certain effects that are desired, and transparency in ASD related applications?

Jacqueline Nadel: For me, as far as possible the best is of course transparency. The parents are aware of what is done during the experiment and when the children have a good cognitive level, they should be aware also. The question has become to be so important that now if the participants are not aware there will be no agreement by the ethics committee. It is really the need of the situation. And then, you can also consider that, at least for high functioning people with ASD, they appreciate honesty; they appreciate people to be simple, to be honest, to be directly asking questions to them, and to be directly explaining what they will do in the situation. Transparency is a need because people with autism

don't know how to lie. They are not liars at all. They don't appreciate it if people are not directly in a situation of honest relation with them. You can try to persuade them that something is good for them providing you really think it is good.

The difference between persuasion and manipulation can be a subtle one and remain implicit in the situation. It is especially true with nonverbal persons. But manipulation can also be a deliberate strategy. I am thinking of social psychology. You can sometimes have participants that do not know what the objective of the research is. They will have been informed about something that is not the real objective. As far as I know at the moment this is really not possible anymore. I know of a lot of research in the field that will not be possible now. They have been done 10 or 15 years ago. Now you will not do this kind of research. But it remains that for people with low performance and people that are nonverbal, the difference between manipulation and persuasion is a very difficult question. What is often observed is that people with autism have real difficulties in choosing, making a decision, taking an initiative. If you start persuading a person with autism that something is good for them, they can be encouraged to do so; and so, persuasion will have a good effect in the situation. But manipulation is something else. You manipulate when you propose a procedure without indicating that you can do something else. You propose how to behave and when to behave. Many instructed programs do manipulate insofar as there is nothing else to do than to follow the instructions. So, the person is not free to say 'no' or to do something else but has to follow the instructions. But here we meet an ethical issue that will prevent manipulation thanks to the investigation of the ethics committee. Preventing the person from being manipulated is really a rule that the ethics committee will have in mind.

### A.6.4 Question 3:

Catherine Pelachaud: You have already answered our third question which is should SIA be better than humans. But 'better than human' is somehow a manipulation.

Jacqueline Nadel: Absolutely, your question, of course, is perfectly valid for human partners, and even maybe more. Sometimes we can measure better if a system is manipulating a child than if a human being is manipulating a child. For instance, if I take the example of turn-taking. Turn-taking is a very good parameter to measure reciprocity and to measure the involvement of the user with the system. You can reorganize the system if you see that the child does not take its turn, that the social agent is always taking the initiative in the situation, always initiating things that the user follows. I think there are many implicit aspects in play in a situation of dependency or addiction. With a social agent you can more easily find the solution than for humans. For instance, there can be a program developed in order to stop positive feedbacks between social agent behavior and the behavior of the person. Imagine the person is asking repeatedly the same response to the

social agent, or the person is imitating repeatedly the behavior of the agent. You can stop these positive feedbacks. You can consider the interactive mode between the social agent and the individual with autism. This is not true for human partners. It is very difficult to stop a stereotypical behavior when it appears during an interaction between a human partner and the child with autism, because it will break not only the interaction but also the relationship.

I think the danger of dependency and addiction is potentially less important for SIA than for human partners. About addiction, you can have an a-priori agreement with the person with autism concerning the duration of use of the SIA for instance. If you have an agreement for a timer, then you will have an interaction with the social agent during 10mn and no more. So, you time the time-timer and everything will stop after 10 min. You will avoid dependency because you will have, at first, limited the time when the user will be in the presence of the social partner. This is something you can do with the social agent that is far more difficult to do with a human partner. Thus with a SIA you are in a better situation than us to manage the question of dependency.

Now for addiction it is something different because it depends largely on the objectives of the program. If the program is a short- term program, there is not much of a problem of addiction. If you have a six-weeks program where each week the child will meet the social agent two times, there is no real danger of addiction. But if it is a long-term program, like for instance you would like the child to develop social skills and the period will be one year or two years, then in this situation, of course, there could be a big problem of dependency and addiction. For long term developmental programs, addiction may appear progressively. I would say that addiction is maybe a special problem with social agents because they are better than humans.

### A.6.5 Question 4:

Birgit Lugrin: This nicely leads us to our next question. You are saying SIAs can be better than humans. So, what does it involve for SIAs to be better than humans?

Jacqueline Nadel: To some extent they are better. They are always ready to welcome, never in a bad mood, never in a hurry to receive an answer, they don't look in the eyes, they are less sophisticated, they are more predictable. These fit particularly well the specificities of people with autism. It is an enormous advantage of social agents compared to humans. So, I would say that these capacities of the animated social agents are a very good way to allow children with autism to accept social situations. A lot of times, social situations are very difficult for people with autism because we are so different from one moment to another. Our eyes are always moving. Our facial expressions are always changing. It is something that makes us unpredictable for children with autism.

All the basic affiliative behaviors will be easier to learn with a SIA rather than directly with a human partner. The virtual partner will have an enormous advantage compared to a human being. Of course, you should progressively decrease these specificities of the social agent in order to make their behaviors more similar to human behaviors. Maybe that is really the problem. You start with a social agent that has these wonderful advantages to be more predictable, to be always in a good mood, to be always ready to welcome; that perfectly fits the specificities of the child with autism. But progressively the child with autism has to adapt to human specificities. So, they have to adapt to unpredictability. That maybe something to think about: how do you make your social agent less and less virtual and, more and more human in the way it interacts with the child? That is the agent becomes less predictable, becomes less happy to meet the child, becomes sometimes joyful, sometimes neutral; and this according to what the child has to understand about the social agent. May be this is the problem. But a good social agent, happy to welcome the child, is the best we can offer to the child with autism at the beginning of a social training. I do not see any danger with that. It is a good way to start a social training. Afterward, the big problem will be about dependency. It is especially an important matter if the social agent takes initiatives and directs at some extent the individual's choices. Then an agreement should always be part of the situation and the SIA should first ask the person to decide what to do and how. For people with no verbal language, the SIA should understand gestures and facial emotional expressions. Also, a very important element is that they should adapt to the personal tempo of the user, the personal rhythm. Some are speedy. But the majority of people with autism, especially nonverbal persons, are slow in their answer. If you don't wait you will take the initiative in the place of the person. Thus, the social agent may interact in a way that takes into account the rhythm, the tempo of the user. You see what I mean.

Birgit Lugrin: Yes, definitively.

### A.6.6 After the questions: Free comment

Jacqueline Nadel: Your clever questions are all part of a solipsistic view of the individual. As soon as we consider the person with ASD as part of a dynamic system relating them to others, all the answers can be modulated by the nature of the dynamics: the real problem of ethics lies in the fact in recognizing the individual with ASD as equal to you in a no-hierarchy conception of human rights.

## A.7 Concluding Remarks

In this challenge discussion chapter, we have seen thought-provoking and critical discussions on the current challenges in research and development of SIAs. These challenges covered both technical challenges and societal or ethical challenges. Although extensive research and

development in the fields of SIAs have drastically advanced the state of the art in the last two decades, there is still a long way to go before we will achieve agents that can truly socially interact whilst being of practical use for people in their intended social domain. This is particularly prominent when we are looking at interactions in the wild and over longer periods of time. But we have also seen that there is lots of room for theoretical research in the lab to completely understand the underlying mechanics of social interaction with artificial entities.

With this handbook, we aimed to bridge the gap between the two communities of IVAs and SRs. We have seen in each chapter of this handbook that there are very common research directions, ideas, challenges and approaches. The challenge discussions particularly highlighted the need for and great benefits of the two communities working together and looking at each other's implementations and research findings.

However, all the works presented in this handbook have also shown that the research conducted by this community is of great interest for (and can largely benefit from input of) other domains such as, for example, virtual/augmented reality, affective computing, game design, computer animation, or Kansei. Studies have demonstrated how users attribute communicative and emotional intention to autonomous agents with abstract figures, not only with human-like appearance; and that those results can apply also to autonomous entities such as voice assistants, conversational agents, assistive robots, but, why not also to autonomous cars. Theoretical and computational models on emotion, cognition, but also behaviours, speech, social space to name some chapters, can be of use not only to model and build embodied agents but also other autonomous entities. The communicative and emotional functions may be common to many of these entities; while their instantiation into behaviours will depend on their embodiment (voice, text, object, etc.) and context of use. These areas are often also located in socially interactive domains, and thus address similar psychological questions as well as technological ones. Furthermore they will also be out for interaction with humans in the wild in the future, and can thus benefit from the research findings presented in this book.

We are thus positive about our endeavour to bring the communities of IVAs and SRs closer together and are inviting other communities to join our journey!

# Bibliography


T. Bickmore. 2022. *Health-Related Applications of Socially Interactive Agents*, pp. 403–435. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application. ACM, 2. DOI: https://doi.org/10.1145/3563659.3563672.

J. Gratch and G. Lucas. 2021. *Rapport Between Humans and Socially Interactive Agents*, pp. 433–462. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM, 1. DOI: http://dx.doi.org/10.1145/3477322.3477335.

B. Lugrin. 2021. *Introduction to Socially Interactive Agents*. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM Press. DOI: http://dx.doi.org/10.1145/3477322.3477324.

B. Lugrin, C. Pelachaud, and D. Traum. 2021. *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*. ACM Press.

A. Paiva, R. Oliveira, F. Santos, and P. Arriaga. 2021. *Empathy and Prosociality in Social Agents*, pp. 385–431. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM, 1. DOI: http://dx.doi.org/10.1145/3477322.3477334.

C. Saund and S. Marsella. 2021. *Gesture Generation*, pp. 213–258. The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition. ACM, 1. DOI: http://dx.doi.org/10.1145/3477322.3477330.