Report from Dagstuhl Seminar 21381

Conversational Agent as Trustworthy Autonomous System (Trust-CA)

Edited by

Effie Lai-Chong Law[1], Asbjørn Følstad[2], Jonathan Grudin[3], and Björn Schuller[4]

1 **Durham University, GB,** lai-chong.law@durham.ac.uk
2 **SINTEF – Oslo, NO,** asbjorn.folstad@sintef.no
3 **Microsoft – Redmond, US,** jgrudin@microsoft.com
4 **Universität Augsburg, DE,** schuller@informatik.uni-augsburg.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 21381 "Conversational Agent as Trustworthy Autonomous System (Trust-CA)". First, we present the abstracts of the talks delivered by the Seminar's attendees. Then we report on the origin and process of our six breakout (working) groups. For each group, we describe its contributors, goals and key questions, key insights, and future research. The themes of the groups were derived from a pre-Seminar survey, which also led to a list of suggested readings for the topic of trust in conversational agents. The list is included in this report for references.

Seminar September 19–24, 2021 – http://www.dagstuhl.de/21381
2012 ACM Subject Classification Human-centered computing → Human computer interaction (HCI)
Keywords and phrases Conversational agents, Trust, Trustworthiness, Autonomous Systems
Digital Object Identifier 10.4230/DagRep.11.8.76

1 Executive Summary

Effie Lai-Chong Law (Durham University, GB)
Asbjørn Følstad (SINTEF – Oslo, NO)
Jonathan Grudin (Microsoft – Redmond, US)
Björn Schuller (Universität Augsburg, DE)

The overall goal of the Dagstuhl Seminar 21381 "Conversational Agent as Trustworthy Autonomous System" (Trust-CA) was to bring together researchers and practitioners, who are currently engaged in diverse communities related to Conversational Agents (CA), to explore challenges in maximising the trustworthiness of and trust in conversational agents as AI-driven autonomous systems – an issue deemed increasingly significant given their widespread uses in every sector of life – and to chart a roadmap for the future conversational agent research. The three main challenges we identified were:

- How do we develop trustworthy conversational agents?
- How do we build people's trust in them?
- How do we optimise human and conversational agent collaboration?

The Seminar Trust-CA took place on 19-24 September 2021 in a hybrid mode. Out of 50 invitees, 19 attended in person and the rest joined online from all over the world, including Brazil, Canada, France, Germany, Greece, Ireland, Netherlands, Norway, Poland, South Korea, Sweden, Switzerland, UK and USA.

The four-day scientific programme started by unpacking the notion of "trust in conversational agent" with a panel discussion. Each of the four seminar organisers expressed their views on the notion. Jonathan Grudin presented a list of ten species of trust that can be applied to conversational agents, for instance, "Trust that a CA will correctly interpret my question or request; will deliver relevant, reliable, useful information." Asbjørn Følstad first presented an overview of the six themes derived from a pre-Seminar survey (details are in Overview of Working Groups) and then described his recent work on the effect of human likeness of a conversational agent on trust. Björn Schuller presented factors influencing trust in humans, such as being reliable, ethical, moral and charismatic, and in conversational agents, such as being explainable, interpretable and transparent. He also discussed how to measure trust reliably and the danger of overtrust. Effie Law discussed the notion of trust with reference to multidisciplinary theory of trust (e.g. psychological, social, historical), beyond the use of questionnaires to evaluate trust, and identifying applications where agents are of high practical value. Some attendees commented on the ideas shared, e.g., the elusiveness of trust.

The scientific programme comprised two major parts – Talks and Breakout Groups. There were altogether 20 talks, covering a range of topics (see Abstracts). Nine of the talks were delivered in person and the rest online. There were six Breakout Groups with each discussing one of the six themes: Group 1 – Scope of Trust in CA; Group 2 – Impact of CA; Group 3 – Ethics of CA; Group 4 – AI and Technical Development; Group 5 – Definition, Conceptualisation and Measurement of Trust; Group 6 – Interaction Design of CA. Group 1, 3 and 4 had one team each whereas Group 2, 5 and 6 had two teams each. To ease collaboration, individual teams were either in-person or online (except for Group 4 which was in hybrid mode). Each group had three two-hour working sessions . In the evening, each group reported progress and invited feedback for shaping subsequent sessions.

The group discussions led to intriguing insights that contributed to addressing the main challenges listed above and stimulated future collaborations (see the Workgroup Reports). Here we highlight one key insight of each group. Group 1 developed a dynamic model of trust with three stages, Build-Maintain-Repair, which evolve over time. Group 2 drafted a code of ethics for trustworthy conversational agents with eight provisions. Group 3 explored the ethics challenge of transparency from the perspective of conversational disclosure. Group 4 called for increased collaboration across research communities and industries to strengthen the technological basis for trust in conversational agents. Group 5 proposed a framework for integrating measurement of trusting beliefs and trusting behaviour. Group 6 analysed several aspects of multimodality to understand their possible effects on trust in conversational agents. Apart from the scientific programme, the Seminar organised several social events, including after-dinner wine and cheese gatherings, hiking in a nearby historic site, and a music event.

Overall, our Dagstuhl Seminar Trust-CA was considered a success. The major outputs were derived from the pre-Seminar survey (six research themes and a recommended reading list), twenty talks, and six multi-session breakout groups. Thanks must go to the enthusiastic involvement of all attendees in analysing various aspects of the burgeoning topic of conversational agents. Of course, the Seminar could only take place with the generosity of *Schloss Dagstuhl – Leibniz Center for Informatics*. The efficiency and friendliness of the scientific and administrative staff of Schloss Dagstuhl was much appreciated by the organisers and all attendees.

2 Table of Contents

3 Overview of Talks

3.1 Chatbots and Voice Assistants from the Perspective of Machine Ethics and Social Robotics

Oliver Bendel (FH Nordwestschweiz – Windisch, CH)

As a discipline, machine ethics examines the possibilities and limits of moral and immoral machines. Social robotics researches and builds robots that interact with, communicate with, are close to, and map features of humans and animals. In doing so, they have a specific use, such as care, support, or entertainment. In his talk, Oliver Bendel outlined the fundamentals of machine ethics and social robotics and presented conversational agents and complementary systems that have emerged from these disciplines. The GOODBOT (2013) is a chatbot that responds morally adequately to problems of the user. The LIEBOT (2016), also a chatbot, can lie systematically, using seven different strategies. The BESTBOT (2018) is a chatbot that recognizes certain problems and conditions of the user with the help of text analysis and facial recognition and reacts morally to them. In 2019, Oliver Bendel and his team developed the MOME (the name stands for "morality menu"). With the help of sliders, you can transfer your moral beliefs to the chatbot MOBO, which then formulates and responds accordingly. The most recent project to date was SPACE THEA (2021), a voice assistant that demonstrates empathy and is designed to accompany astronauts to Mars. Most of the artifacts earn our trust by recognizing our situation and helping and supporting us. The LIEBOT, on the other hand, systematically lies to us and makes us aware that conversational agents can be designed in an abusive or negative way.

3.2 Interaction with multi-bots

Heloisa Candello (IBM Research – Sao Paulo, BR)

User Evaluation of Multi-party Conversational Systems. Recent advances in artificial intelligence, natural language processing, and mobile computing, together with the rising popularity of chat and messaging environments, have enabled a boom in the deployment of interactive systems based on conversation and dialogue. This talk explores the design and evaluation of conversational interfaces, and it is focused on design and evaluation methods that address specific challenges of interfaces based on multi-party dialogue. I will show two projects. First, Café com os Santiagos is an artwork where visitors conversed with three chatbots portraying characters from a book in a scenographic space recreating a 19th-century coffee table. It was accessed by more than 10.000 users in a public space, resulting in insights to improve the conversation system even more. Second, I will show an experiment with Finch's cognitive investment adviser. Finch interface made a state-of-art artificial conversational governance system accessible for regular users to assist in financial decisions.

3.3 Why should we care about linguistic register? Insights on chatbot language design

Ana Paula Chaves (Federal University of Technology – Paraná, BR)

This talk discusses the relevance of linguistic register as a theoretical framework for chatbot language design. I presented the concept of register and discussed how using register-specific language influences the user's perceptions of their interactions with chatbots. To demonstrate that, I presented a study performed in the context of tourism information search chatbots, where participants evaluated the language appropriateness, credibility, and user experience when facing chatbot utterances in different registers. I argued that the appropriate use of language is relevant to design trustworthy chatbots since it influenced the chatbot's credibility in our study. I also pointed to future research directions

3.4 Towards Personalized Explainable AI

Cristina Conati (University of British Columbia – Vancouver, CA)

The AI community is increasingly interested in understanding how to build artifacts that are accepted and trusted by their users in addition to performing useful tasks. It is undeniable that explainability can be an important factor for acceptance and trust. However, there is still limited understanding of the actual relationship between explainability, acceptance, and trust and which factors might impact this relationship. In this talk, I argue that one such factor relates to the user's individual differences in terms of both long-term, stable traits (e.g., expertise, cognitive abilities, preferences) and short-term transient states (e.g., level of cognitive load, affective state). Namely, given a specific AI application, different types and forms of explanations may work best for different users, and even for the same user at different times, depending to some extent on both their long-term traits and short-term states. As such, our long-term goal is to develop personalized XAI tools that adapt dynamically to the user's needs by taking relevant user factors into account. In this talk, I focus on research investigating the impact of long-term traits, and how they may drive XAI personalization. I present a general methodology to address these questions, followed by an example of how it was applied to ascertain which long-term traits are relevant for personalizing explanations in an intelligent tutoring system (ITS). I discuss how to move forward from these insights, and present research paths that should be explored to make personalized XAI happen.

3.5 In human-likeness we trust? The implications of human-like design on partner models and user behaviour

Benjamin Cowan (University College – Dublin, IE)

In human-likeness we trust? The implications of human-like design on partner models and user behaviour" Abstract: Voice has now become a mainstream interaction modality. Current voice interfaces fundamentally rely on human conversation as an interaction metaphor, using human-like design to support partner model building. My talk will explore how human-like VUI design shapes our beliefs of a machine partner's abilities, how this is potentially crucial to consider in terms of trust in voice interface interaction, and whether this interaction metaphor is actually appropriate as we strive for more trustworthy conversational systems.

3.6 Underestimated Challenges in Developing and Using Conversational Agents

Jonathan Grudin (Microsoft – Redmond, US)

For decades, conversational agents were developed in small, trusting, homogeneous laboratories. Since 1995, commercial internet activity and the web has seen the rise of "bad actors" and a range of grey activity, creating challenges that need to be anticipated by university researchers and conversational agent developers who still work in small, trusting groups where consideration of potential technology misuse is low. The 'virtual companion' artificial general intelligence, reflected in ELIZA and Turing Test contestants, remains a science fiction mainstay but is approaching real-life extinction. The take-down of Tay by trolls led to sophisticated risk-mitigation approaches, but it is an expensive arms race. Amnesic conversational partners are unappealing but privacy considerations inhibit the retention of personal communication. Hugging Face, Zo, Le Luda, Replika, and others disappeared or failed to gain traction. At the brief-conversation extreme, intelligent assistants such as Alexa have encountered concerns about re-enforcing submissive female stereotypes and shaping children to converse with impersonal imperatives. Most work on conversational agents lies between these two. Task-focused chatbot technology can be employed by a range of people with a range of intentions. Facilitating seamless human-in-the-loop can be necessary and wonderful, but concealing that humans are in the loop can be ruinous. Major uses of chatbots include reducing human conversation or more effectively steering behavior. This can yield positive outcomes or negative outcomes. Let's aim high, but periodically consider unintended consequences should our work be misused.

3.7 Designing Conversational Agents for Dyadic and Group Interactions

Soomin Kim (Seoul National University, KR)

The advancements in technology shift the paradigm of how individuals communicate and collaborate. Machines play an active role in human communication. However, we still lack a generalized understanding of how exactly to design effective machine-driven communication and discussion systems. In this paper, I present new interactive systems in the form of a conversational agent, or a chatbot, that facilitate dyadic and group interactions. Specifically, I focus on: 1) a conversational agent to engage users in dyadic communication, 2) a chatbot called GroupfeedBot that facilitates daily social group discussion, 3) a chatbot called DebateBot that enables deliberative discussion. The findings of this thesis are as follows. For a dyadic interaction, participants interacting with a chatbot system were more engaged as compared to those with a static web system. However, the conversational agent leads to better user engagement only when the messages apply a friendly, human-like conversational style. These results imply that the chatbot interface itself is not quite sufficient for the purpose of conveying conversational interactivity. Messages should also be carefully designed to convey such. Unlike dyadic interactions, which focus on message characteristics, other elements of the interaction should be considered when designing agents for group communication. In terms of messages, it is important to synthesize and organize the information given that countless messages are exchanged simultaneously. In terms of relationship dynamics, rather than developing a rapport with a single user, it is essential to understand and facilitate the dynamics of the group as a whole

3.8 Measuring Understanding in Interactions with Embodied Conversational Interfaces: Theory, Studies and Computation

Dimosthenis Kontogiorgos (KTH Royal Institute of Technology – Stockholm, SE)

Main reference Dimosthenis Kontogiorgos, André Pereira, Joakim Gustafson: "Grounding behaviours with conversational interfaces: effects of embodiment and failures", J. Multimodal User Interfaces, Vol. 15(2), pp. 239–254, 2021.
URL https://doi.org/10.1007/s12193-021-00366-y

Research in face-to-face human-robot interaction has focused on developing teaching robots that have little abilities to adapt to users' signals of understanding. Human speakers seem to establish common ground incrementally, the mutual belief of understanding among the conversational partners. When teaching each other new tasks, speakers tend to package pieces of information in small fragments and provide information to the learners incrementally. In this talk, I will present our work investigating how speakers' incremental construction of utterances affect the cognitive resources of the conversational partners during utterance production and comprehension. I will also discuss implications for future empirical research on the design of task-oriented human-robot interactions, and how assistive social robots may benefit from the production of fragmented instructions. Using data from a recent online perception study, I will finally present empirical findings from recent research on how we used mouse movement analysis to detect user uncertainty when guided by a conversational interface.

3.9 Establishing long-term relationships with conversational agents – lessons from prolonged interactions with social robots

Guy Laban (University of Glasgow, GB)

Social robots' cognitive architectures and embodied cognition can elicit socially meaningful behaviours and emotions from humans. These robots can afford valuable opportunities for social engagement with human users, and there is a growing evidence base that documents how social robots might function as autonomous tools to support psychosocial health interventions via establishing meaningful relationships. Since interactions with social robots are novel and exciting for many people, one particular concern is the extent to which people's behavioural and emotional engagement with robots might develop from initial interactions with a robot, when a robot's novelty is especially salient, and be sustained over time. Here we aimed to test the extent to which social robots can elicit emotional expression and disclosures from people, as well as affect their perceptions in a long and intensive period. Through the use of a mediated online experiment, this research was designed to examine the type and extent of expressions people use to communicate with a social robot, how they perceive it, as well as how people disclose information and emotions to a social robot via online video chats across time. Across a period of five weeks, 39 participants engaged in interactions with the social robot Pepper (SoftBank Robotics) via Zoom video chats twice a week. Participants were asked by Pepper about their general everyday experiences in one condition, whereas in the second condition these topics were framed to the COVID-19 pandemic. Our results suggest that people gradually perceived the robot to demonstrate higher degrees of agency and experience across sessions, as well as being friendlier. Moreover, participants perceived the interaction quality and the robot communication competence to be better across sessions. Finally, participants reported positive mood changes due to their interactions with Pepper across all sessions.

3.10 A Cryptocurrency Chatbot and the Social-technical Gap of Trust

Minha Lee (TU Eindhoven, NL)

Main reference Minha Lee, Lily Frank, Wijnand A. IJsselsteijn: "Brokerbot: A Cryptocurrency Chatbot in the Social-technical Gap of Trust", Comput. Support. Cooperative Work., Vol. 30(1), pp. 79–117, 2021.
URL https://doi.org/10.1007/s10606-021-09392-6

Cryptocurrencies are proliferating as instantiations of blockchain, which is a transparent, distributed ledger technology for validating transactions. Blockchain is thus said to embed trust in its technical design. Yet, blockchain's technical promise of trust is not fulfilled when applied to the cryptocurrency ecosystem due to many social challenges stakeholders experience. By investigating a cryptocurrency chatbot (Brokerbot) that distributed information on cryptocurrency news and investments, we explored social tensions of trust between stakeholders, namely the bot's developers, users, and the bot itself. We found that trust in Brokerbot and in the cryptocurrency ecosystem are two conjoined, but separate challenges that users and developers approached in different ways. We discuss the challenging, dual-role of a Brokerbot as an object of trust as a chatbot while simultaneously being a mediator of

trust in cryptocurrency, which exposes the social-technical gap of trust. Lastly, we elaborate on trust as a negotiated social process that people shape and are shaped by through emerging ecologies of interlinked technologies like blockchain and conversational interfaces

3.11 Codex as a personal assistant?

Clayton Lewis (University of Colorado – Boulder, US)

The development of language-model based artificial intelligence, as seen in Codex and GPT-3, may open the way to new forms of artificial personal assistants. At present, a trial of Codex produces interesting results for a use case involving keeping track of gifts. The generated code does not work, and would require programming knowledge to repair. But the results suggestion that these models may offer new ways to think about the challenges of cognitive science, including the challenges to cognitive theorizing articulated by Harold Garfinkel. These ways of thinking may also contribute to understanding the mechanisms of trust in interactions with artificial agents.

3.12 Designing Inclusive Conversational Agents that Older Adults Can Trust

Cosmin Munteanu (University of Toronto Mississauga, CA)

Older adults (65+) are at increasing risk of being "digitally marginalized" or "digitally isolated". This is often the result of active or passive, conscious or unconscious bias in how older adults are overlooked in the design of new digital technologies. When design is more actively focused on older adults, this is often reduced to clichés about limited cognitive or physical abilities. The consequences of such approaches are significant, with many seniors having difficulties in transitioning their use of essential services to the online space in several key areas: taking financial decisions, understanding health information or accessing health services, staying connected to families, or simply doing online shopping. This is, paradoxically, exacerbated by the increased use of interfaces that are marketed as "natural", such as voice and conversational agents (chatbots). In this talk I focus on one of the most overlooked barriers toward older adults' trusting of such interfaces: mental models. I am arguing for new methodological approaches that empower older users and put them in the lead for designing novel interactive technologies that assist with reducing their digital marginalization and isolation, and through this, better reflect older adults' mental models of interacting with and trusting of such new technologies.

3.13 An introduction of the UKRI Trustworthy Autonomous Systems Node on Trust

Birthe Nesset (Heriot-Watt University – Edinburgh, GB)

Robots are rapidly gaining acceptance in recent times, where the general public, industry and researchers are starting to understand the utility of robots, for example for delivery to homes or in hospitals. However, it is key to understand how to instil the appropriate amount of trust in the user. One aspect of a trustworthy system is its ability to explain actions and be transparent, especially in the face of potentially serious errors. Here, we study the various aspects of transparency of interaction and its effect in a scenario where a robot is performing triage when a suspected Covid-19 patient arrives at a hospital. Our findings consolidate prior work showing a main effect of robot errors on trust, but also showing that this is dependent on the level of transparency. Furthermore, our findings indicate that high interaction transparency leads to participants making better informed decisions on their health based on their interaction. Such findings on transparency could inform interaction design and thus lead to greater adoption of robots in key areas, such as health and well-being.

3.14 Impact of adaptation mechanisms on user's perception of agent

Catherine Pelachaud (Sorbonne University – Paris, FR)

During an interaction, interlocutors may adapt their behaviors at different levels. We have developed different adaptation mechanisms for a virtual agent interacting with human users. These mechanisms act at the levels of conversational strategy and multimodal behaviors or at the signals level. The former two mechanisms are modeled using reinforcement learning technique. The agent learned the best strategy or multimodal behaviors to display on the fly while interacting with a user. The third model is learned from data of human-human interaction and is modeled using LSTM. These three mechanisms have been integrated within an architecture for human-agent interaction. Lately we have worked on integrating a modality that has not received much attention in human-agent interaction, namely touch. Social touch conveys several functions such as showing emotion, getting the attention, comforting someone. We have developed a decision model, based on the emotional model FAtiMA to endow the agent with the capacity to determine when to touch the user and with which touch. Finally, a third argument I have presented regards our work on simulating laughter in virtual agent, and in particular, how laughter can have an impact on users perception of the agent and on the quality of the interaction.

3.15 To trust or not to trust? What is the use case? Insights from applied research in conversational interaction

Stefan Schaffer (DFKI – Berlin, DE)

Applied research at DFKI provides insights into various application examples of conversational interaction. Three demonstrators for conversational assistant system projects for the use cases of care management, railway security and museum guides are presented and references to the topic of trust are made. The main requirement in the care management use case was to improve the inclusion of visually impaired caregivers. Based on user research, a demonstrator of a conversational ERP tool for care management was developed. A demonstration of the system showed that correctness, which is affected by automatic speech recognition errors, for example, is a key factor in ensuring trust. The security service use case focusses on the issue that soccer fans in rail travel often cause security relevant situations. Based on participatory design, a conversational assistant for efficient input of security relevant information was developed. A focus group revealed that generation of reliable information is crucial for this use case. Regarding trust, the recommendation was derived that appropriate system feedback should be generated for safety-relevant information. In the museum use case, several NLP mechanisms, including the transformer-based model BERT, are implemented to answer fact and open questions. To make the necessary annotation effort manageable, different amounts of training data are annotated for different objects. The assumption is stated that the total amount of meta data enrichment will influence the level of trustworthiness of the system.

3.16 The value of small talk and responsiveness

Ryan Schuetzler (Brigham Young University – Provo, US)

Small changes in the way a CA interacts can influence both behavior and attitudes. Tailoring chatbot messages to reflect active listening demonstrates to users that the bot can understand them, which can improve feelings of social presence and engagement. In an interview chatbot, we manipulated tailoring in the small talk rapport-building phase of the interview to understand the effect it has on perceptions and self-disclosure. In most circumstances, increased social presence and engagement is a good thing. However, we have shown that in discussions of sensitive information, less social presence might be preferable to improve user disclosure. Because users disclosing sensitive information want to feel that they are not being judged, tailoring reduces disclosure. While tailoring and small talk are small manipulations, they can create significant effects in how users perceive and respond to a chatbot.

3.17 How and When should chatbot self-disclose?

Zhou Yu (Columbia University – New York, US)

Social chatbots research has attached much attention lately. However, how and to what extent people respond to chatbot self-disclosure and how self-disclosure can impact task success remain less known. We designed a social chatbot that can perform three different types of self-disclosure: sharing factual information, cognitive opinions, and emotions. The chatbot can conduct small talks and provide relevant recommendations on two topics, movies and COVID-19 best practices. Through a large-scale user study, we found that chatbots' level of self-disclosure correlates with better conversational engagement and warmth towards the chatbot. Chatbots that perform all three types of disclosure also complete the recommendation task more effectively than ones that only perform one or two types of disclosure.

3.18 Democratizing Conversational AI: Challenges and Opportunities of No-Code, Reusable AI

Michelle X. Zhou (Juji Inc. – Saratoga, US)

Creating quality conversational AI agents not only requires deep AI expertise and sophisticated software engineering skills, but also requires large amounts of training data and intensive computational resources. Few organizations have such expertise, let alone the required resources to develop and manage their own version of conversational AI agents. To democratize conversational AI and bridge the potential AI divide, we have been developing an end-to-end, no-code AI platform that enables non-IT professionals to create, deploy, and manage their custom conversational AI agents with no code, and no IT resources required. Such a conversational AI platform has three key characteristics. First, it supports the end-to-end, no-code development of conversational AI agents with cognitive intelligence–AI agents with human soft skills, such as active listening skills and reading between the lines. These human soft skills enable AI agents to interact with their users and complete their tasks responsibly and empathetically. Second, it supports multi-level reuses of pre-built AI components, which then enables rapid customization of a conversational AI agent with no code. Third, it enables real-time conversational AI monitoring and live updates/improvements without interrupting ongoing critical conversations. Our platform has been used by non-IT professionals from multiple domains to create and manage their own conversational AI agents, demonstrating the practical values of no-code, reusable AI.

4 Working groups

4.1 Overview of Working Groups: Origin and Process

All Groups

4.1.1 Origin

Prior to the launch of the Seminar, a web-based survey was conducted to gather the attendees' views on the following issues:

1. What are the main challenges to be addressed for the topic of trust in CA?
2. Which papers (max. three) to be recommended as key background reading for the topic of trust in CA?
3. What topic to be proposed for a PhD student research project related to trust in CA?

Responses from 39 attendees were obtained. Thematic analysis of the data for *Item 1* resulted in six themes as shown in the concept map below. Each of these themes was discussed in a breakout group during the Seminar (Section 4.2 – 4.7). For *Item 2*, a list of references was compiled (Section: Open Problems). For *Item 3*, only a subset of the respondents provided input, results are not presented here.

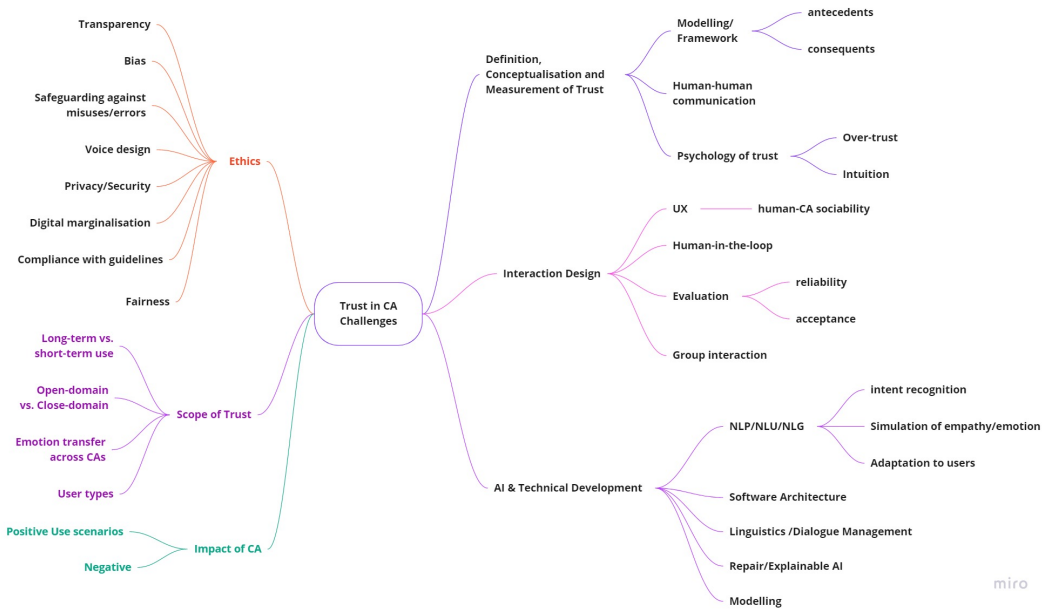

Figure 1 Six themes of Trust in CA.

4.1.2 Process

The aim of conducting breakout groups was to advance state of the art theories, methodologies and practices on trust in CA. The group discussion, which was guided by some key questions, drew on the experiences and expertise of individual members. Each of the breakout groups

had three sessions with each lasting about two hours. Key insights from the discussion were reported back in a plenary meeting in the evening to invite feedback for shaping the direction of the groupwork in the following day.

The 50 attendees were allocated to different breakout groups based on their preferences. An attendee joined one group in the morning and another group in the afternoon. Different group memberships as such encouraged stimulation and collaboration. Outputs of each breakout group are summarised in the following subsections.

4.2 Breakout Group 1: Scope of Trust in CA

Group1

Contributors: Oliver Bendel, Birthe Nesset, Catherine Pelachaud, Guy Laban, Eren Yildiz, Effie Law

4.2.1 Goal and key questions

Goal: To further explore the theoretical and practical basis for trust in CAs.
Key questions:
- How is trust in CAs established and maintained?
- Which are the relevant factors?

Relevant aspects: Usage duration (long-term vs. short term use); Domain specificity (open domain vs. closed domain); Transferability of experience across different CAs; User types

4.2.2 Key insights

A dynamic stage-based trust model with the temporal aspect is the key insight gained from the discussions of Group 1. Specifically, there are three main stages of trust evolvement, namely Build, Maintain and Repair. Depending on the severity of the consequence of broken trust between a CA and its user, interaction strategies deployed in a preceding stage may be invoked.

The Build stage. A chatbot starts an interaction by building trust between the user and itself. It is crucial that the chatbot knows whether trust is established, and that the user is aware of the scope of the chatbot. To build trust, the following factors should be taken into consideration:
- showing affordances such as abilities, core features, limitations, and purposes
- setting up the right expectations and cost
- personalization and customization
- giving a sense of control to the user
- accommodating the user's mistakes
- recognizing the user's goals, needs and preferences and taking them seriously

When an optimal level of trust is established, the chatbot can switch to the Maintain Stage.

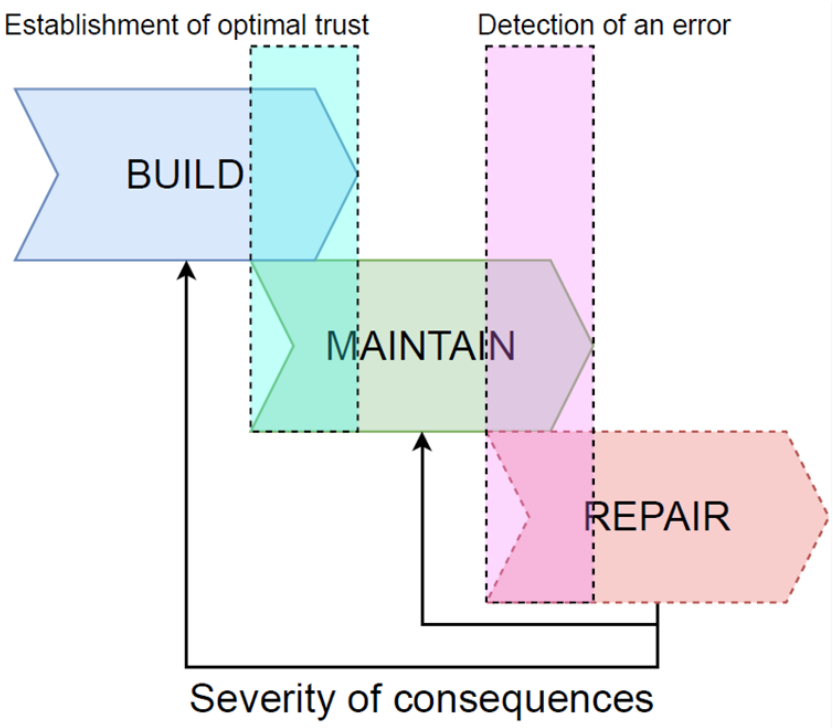

Figure 2 Dynamic Trust Model.

The Maintain stage. It is crucial to maintain the relationship between the chatbot and its user to ensure the continuity of user engagement. The chatbot can achieve so with the following features:

- being adaptable and reliable
- obtaining ongoing feedback from users to analyse their behaviour, attitude and emotions.
- showing the chatbot's ability to recognise the user with reference to the interaction history

The Maintain stage is the desired stage of the chatbot for demonstrating its trustworthiness. However, when some components of the chatbot fail to accomplish their task, it moves to the Repair Stage.

The Repair Stage. The chatbot aims to fix trust issues between the user and itself. The following actions are required:

- acknowledgment of error
- identification of error
- apologising
- repair and correction by learning
- reaffirming conforming to the shared goal

However, some errors do not need to be repaired and the interaction can be transferred back to the Maintain stage. The chatbot may decide that a user action, which has broken trust due to the failure of a certain task, is not important. Thus, it is acceptable to continue having trust in that component as it is. On the other hand, if the user detects an error committed by a chatbot and points it out, the chatbot can attempt to repair trust by issuing

an apology. Another condition is that if the chatbot is not confident about the accuracy of a response but delivers it nevertheless with a forewarning, it could mitigate the need to repair. In some cases where the chatbot decides that the broken trust cannot be repaired because it is either too costly to perform the repair or the repair is impossible with the approach used, then the chatbot may decide to build a new relationship for the same goal but with a different approach.

Overall, by assessing the severity of consequences of the broken trust, the chatbot decides whether (i) to repair trust, (ii) switch to the stage of maintaining the relationship, or (iii) build a completely new relationship for the same goal with a different strategy/approach.

Regaining Trust. The following actions can be undertaken to regain the lost trust:
- adjusting the weights of individual factors of trust; such weights are application-dependent and user-specific
- referring to a taxonomy of CA can help fine-tuning the weights, which can also be supported by participatory design and empirical evaluation
- real-time signal detection and adaptation to allow CA to clarify intents, manage user expectations and update user models as strategies to adjust the weights of CA trust
- resolving mismatch between error performance and mental models (e.g., user verbal and non-verbal behaviours to infer emotion)

Basically, every chatbot type has different weights for individual trust factors, including:
- Inclusiveness (e.g., accessibility, non-discriminativeness)
- Competences
- Availability
- Warmth (e.g., friendliness, empathy)
- Legality
- Engagement
- Reliability (i.e., consistency)
- Professionalism (e.g., type of the language, avoiding typos and grammar mistakes, embodiment appearance)

4.2.3 Future Research

The following questions require further research effort to address:
- The notion of 'modality' entails clarification: Is multimodality integral to CAs or an add-on feature?
- How to define and operationalise the features (the above bullet points) in each of the three stages of the dynamic and temporal trust model?
- Are factors of trust hierarchical? Whether and how they can be prioritised?
- How can machine learning methods be used to determine the weights of individual factors, and adapt them with respect to contextual changes?

4.3 Breakout Group 2: Impact of CA

Group 2

Contributors:
Group 2a (in person): Sebastian Hobert, Ryan Schuetzler, Frode Guribye, Clayton Lewis, Stefan Schaffer, Martin Porcheron
Group 2b (online): Levi Witbaard, Margot van der Goot, Stefan Morana, Ana Paula Chaves, Jonathan Grudin, Heloisa Candello, Christine Liebrecht, Yi-Chieh Lee, Jasper Feine

4.3.1 Goal and key questions

Goal: Trust in CAs through positive social and commercial changes.
Key questions: How may CAs be applied for positive social and commercial impact?
Relevant aspects: Positive use scenarios; Negative use scenarios

4.3.2 Key insights

4.3.2.1 Group 2a

The group started with the basic question "What is social good?". Then they explored the ethical implications of designing CAs, identified research and development areas of CA as well as future work.

- **Definition of Social Good**: It can be defined in many forms, but in our breakout group, we adopted a utilitarian demarcation in which you design with the aim to maximize benefits for the individual users and society at large, while minimizing individual and societal risks. Of course, one can never predict the full consequences of any action, nor can we predict all the ways people will adapt to the affordances provided by our systems. Intent is core to driving a project for social good; designers, developers, organisations must adopt a stance to delivering social good.
- **Draft for a Code of Ethics for trustworthy CAs**: If driving social good is rooted in the design of CAs, as per our definition, we propose the following set of ethical guidelines for designers to consider:
  - Design CAs and their underlying systems to be worthy of users' trust, not just with the appearance of trustworthiness
  - Be open and explicit about the intent of the CA
  - Take care to recognize and design for marginalized groups
  - Consider the possible negative uses of the CA
  - Take responsibility for the intended and unintended consequences of CA use
  - Minimize risk of harm created by inaccurate responses, or through disclosure of private information
  - Consider possible sources of bias, including commercial interests, and be explicit when they might conflict with the users' best interests
  - Do not unnecessarily exploit the humanlike capabilities of a CA to deceive or manipulate users

- **Research and Development Areas for Social Good**: We believe that the strongest possibilities for social good chatbots are present when one or more of the following conditions are met: (1) the humanlike conversational capabilities of a CA that allow it to accomplish its goal better than a traditional system; (2) the CA can do things that people either cannot (either through lack of ability or resources) or are not willing to do. We have identified the following categories for the creation of CAs for social good. For each category, we begin to outline promising areas for development as well as potential pitfalls that must be addressed to maximize social good and minimise the risk of harm.
  - **Mental Health**: CAs can relieve shortages of mental healthcare workers and reach those who may not otherwise have access or may otherwise be hesitant to seek access. The digital nature of CAs can enhance user trust, and especially their trust that they are not being judged or evaluated based on their responses. However, especially in this category of CA, care must be taken to minimize the potential for inaccurate responses causing harm. While a CA can potentially reach and help many people, designers and researchers must do all they can to ensure the appropriate response to crisis situations such as suicidal thoughts.
  - **Virtual Companionship**: Loneliness has become epidemic, across all age groups under different conditions. Our CAs have the potential to relieve loneliness and provide a connection to those that may otherwise not have one. These agents can be designed to help the elderly who tend to experience loneliness more than others. In the design of these virtual companions, designers must take care to ensure individuals do not develop a dependency on the technology and avoid otherwise beneficial human contact in favour of virtual companionship.
  - **Learning**: The main purpose of CAs in educational settings is supporting instructors, teaching assistants or learners in-class, blended- or online settings instead of replacing them. Providing (automated and individualized) feedback, learning materials, or answers to individual questions in a conversational way seems to be useful particularly useful in large-scale settings in which otherwise learning support would not be offered or in small-scale settings in which offering manual support is effortful. This seems valuable to social good as a more educated populace worldwide increases. Some people especially benefit from a social connection associated with learning, as evidenced by the struggles some experienced during the virtual learning of the COVID-19 pandemic.
  - **Healthcare**: With a shortage of healthcare workers worldwide and limited resources to provide appropriate care, CAs have the potential to relieve pressure on strained human resources. By providing automated access to, for instance, screening and informational services, we can enhance availability and access to these vital services. CAs can also reach populations currently unserved or underserved by the healthcare system.

- **Participatory Design Activities to Ensure Trust in CA**: Since there might be a variation in perspectives, preferences, goals and values in terms of how social good is perceived as such we should not rely solely on the perspectives of the designers and developers when creating CAs. One way of ensuring that the design of a CA is properly anchored in the perspectives and values of its future users is to include them in design activities. Providing design environments and authoring tools that are easy to use can be an important step in empowering communities of practice to build their own CAs that meet the particular needs of the community. With available authoring tools for CAs, participants can engage in design activities that are more similar to end-user programming and tailoring of services and skills that can be aimed at meeting such particular needs.

- **Considerations of Trust**: Trust is especially important in "social good" applications because they often, but not always, deal with higher levels of risk. Trust, and trustworthiness, are important determinants of use when risk is high. As social impact applications are often used with emerging markets and unserved or underserved populations, trust and trustworthiness are critically important to not:
  - (Premature) Deployments of ineffective or harmful technology can hamper future research and developments, slowing down new developments for years to come (e.g., Clippy, Tay, Google Glass, or an over-eager deployment of self-driving cars). Over-promising and setting too-high expectations could erode future trust if the technology fails to meet expectations, even if it is better than alternatives. (e.g., even if a self-driving car is better than human drivers, it faces increased scrutiny, and if it kills people, it erodes societal trust and hinders the advance of future, better technology).
  - Because CAs are still somewhat an emerging technology, malicious, ineffective, or harmful use could result in erosion of trust at a general level (e.g., if Amazon was found to be selling information from private conversations near Alexa)

4.3.2.2 Group 2b

Outcomes of the discussion are summarised in Figure 3:

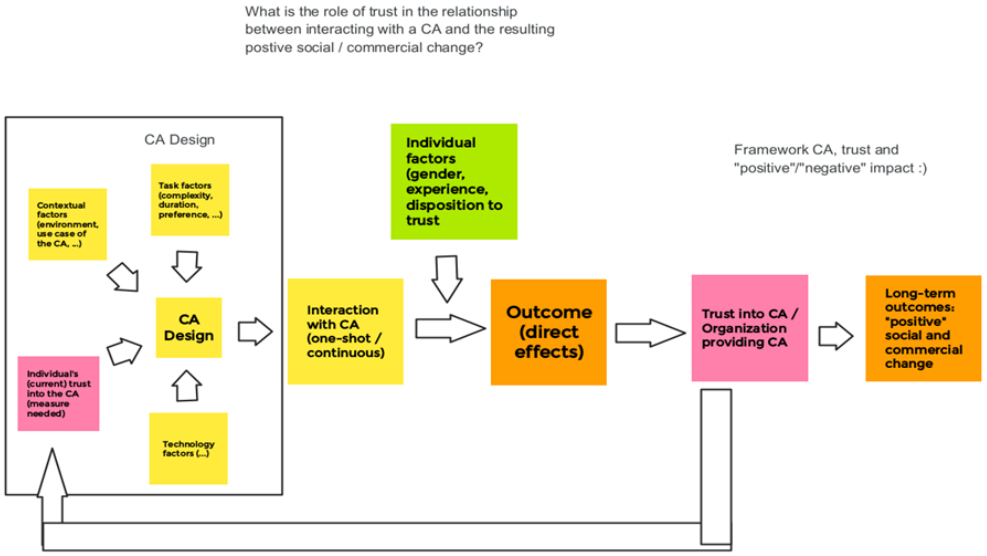

Figure 3 CA design and impact.

Accordingly, CA design is influenced by the contextual, technological and task factors. Interaction with CA (one-shot or continuous) leads to outcome in terms of direct effects, including trust in CA itself and the organisation providing CA.

4.3.3 Future Research

The following research questions need to be explored as future work:
Group 2a:

- Do transparency and explanations support trust in CAs for social good?
- Effectiveness – do our social good CAs actually produce social good, producing better outcomes for all or certain individuals or groups
- How to foster initial trust in CAs (for social good) and how to maintain the impression of trust over time?
- Trustworthiness vs. impression of trustworthiness? What is more important?
- Intent seems to be a core concept. Is intent the only difference related to trust between commercial CAs and CAs for social good?
- How to manage expectations in a new market without previous CA experience?

Group 2b:

- Should the CA remember? Should the CA immediately show that the CA has the history knowledge? To what extent is longer-term interaction needed from the perspective of the user?
- Should CA address stereotypes (e.g. gender stereotype) and how we as a community can contribute to fight against these stereotypes?
- How to continuously/automatically measure trust in CA? Are there approximations for trust rather than using a survey?
- Should the CA be able to measure user trust and adapt itself depending on the user's current level of trust?
- Continuous real-life long-term relationship (e.g. financial trading) is more difficult to study in an experimental context. How to ensure match between real-world case study and study design?

4.4 Breakout Group 3: Ethics of CA

Group 3

Contributors: Minha Lee, Björn Schuller, Elisabeth André, Leigh Clark, Asbjørn Følstad

4.4.1 Goal and key questions

Goal: Trust in CAs through ethical design and implementation.
Key questions:

- Which are key ethical aspects of CAs?
- How to design for ethical CAs use?

Relevant aspects: Transparency, bias, fairness, and digital marginalization, privacy/security, safeguard against misuse and error.

4.4.2 Key insights

Initial considerations – scoping the ethics challenge. Ethics is a research topic of high relevance to trust in conversational agents. In their overview of future directions in chatbot research, Følstad et al. (2021) identified ethics as an area in need of substantial research efforts. Such future research has a valuable starting point in the existing background on ethics in AI-based systems (Hagendorff, 2020). For example, an EC High level Expert Group has detailed the ethical basis for trustworthy AI-systems in general (EC, 2019). Research on ethical aspects on conversational agents is also emerging (e.g., Ruane et al., 2019).

Given the broadness of research challenges pertaining to ethics in conversational agents, the group converged on a specific research challenge of high importance for ethics in conversational agents: conversational disclosure.

Conversational disclosure – a key ethical challenge in CAs. Conversational disclosure concerns how to achieve transparency during interaction with conversational agents. Transparency is a key ethical requirement in AI-based systems (EC, 2019; Hagendorff, 2020), and concerns the need to (a) clarify to users that they are interacting with an automated system, not a human, (b) provide insight into how user data are processed and used, and (c) explain the relevant system characteristics and limitations to the user.

Transparency may be a particularly important ethics requirement for conversational agents as the interaction with such agents may easily be confused with interaction with humans, and users may also be inclined to share personal information as part of such interaction – for example as part of using conversational agents for mental health or relational purposes. The forthcoming European legal regulation of AI systems, the AI Act, will likely make transparency in conversational agents a legal requirement (Schaake, 2021) – so that providers are responsible for users understanding they are interacting with a conversational agent and not a human being.

Designing for transparency through conversational disclosure may be achieved by following two different paths: a guidelines-oriented approach and a practice-oriented approach; the two paths corresponding to a deontological vs. virtue-oriented approach to ethics in AI-systems (Hagendorff, 2020). The two paths are detailed below.

Guidelines-oriented approach to conversational disclosure. In a guidelines-oriented approach, it may be valuable to consider how to provide conversational disclosure at different points in time during a prolonged period of use.

- *initial disclosure*, at the onset of the first interaction
- *routine disclosure*, at predefined milestones
- *requested disclosure*, initiated by the user.

For each of these forms of disclosure, research may address which items to include as part of the disclosure and how to design such disclosure so as to provide a good user experience. Could, for example, routine disclosure be designed so as to provide added value to the user? (e.g. presenting content from previous interactions for evocation or engagement, inspired by approaches to sharing insight based on the users' person information in services like Google Timeline or Strava.).

Practice-oriented approach to conversational disclosure. A practice-oriented approach to conversational disclosure would concern establishing conversational disclosure as a craft skill. Establishing such a skill would include tackling challenges such as how to provide conversational disclosure without disturbing the flow of interaction.

For example, users of relational conversational agents may wish for such agents to be humanlike in their interaction so as to achieve a desired perceived companionship. At the same time, such agents should also be fully transparent to their users. Nevertheless, the interaction should not be interrupted at inappropriate points in time by having the agent explain itself.

Negotiating the need for human likeness on the one hand and conversational disclosure on the other is a design challenge that may require craft skill rather than guideline adherence. To establish a practice-oriented approach to conversational disclosure, ethics may need to be included in teaching and training on design of conversational agents.

4.4.3 Future Research

Ethics in conversational agents is an area of a broad range of research challenges. In this groupwork we addressed one such challenge, conversational disclosure. To further explore ethics in conversational agents, the groupwork participants have initiated a forthcoming CHI workshop on ethics in conversational user interfaces.

References

1 Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and Machines, 30(1), 99-120.
2 EC (2019). Ethics Guidelines for Trustworthy Artificial Intelligence. Technical report, EC High level Expert Group on AI.
3 Ruane, E., Birhane, A., & Ventresque, A. (2019, December). Conversational AI: Social and Ethical Considerations. In AICS (pp. 104-115).
4 Schaake M (2021). The European commission's artificial intelligence act. Policy Brief, Stanford HAI

4.5 Breakout Group 4: AI and Technical Development for CA

Group 4

Contributors: Matthias Kraus, Roger K. Moore, Ricardo Usbeck, Ana Paiva, Rolf Pfister, Elayne Ruane.

4.5.1 Goal and key questions

Goal: Towards advancing the technical basis for trust in CAs
Key question: How may the technical basis for CAs be advanced to strengthen trust?
Relevant aspects: NLP – intent recognition, simulation of empathy, adaptation to users, software architecture, linguistics / dialogue management, repair, explainable AI, modelling.

4.5.2 Key insights

The topic of the technological basis for trust in CAs is very broad. The group addressed this by discussing some key issues over the course of the workgroup sessions. Summaries of the discussions are provided below.

The need for strengthened collaboration across research communities. Advancing the technical basis for trust in CAs is challenging due to disconnected nature of research communities. As a starting point for strengthen connections and exchange between communities, a list of research communities was compiled:

- CHI – ACM CHI Virtual Conference on Human Factors in Computing Systems
- CONVERSATIONS – workshop on chatbot research
- CUI – conference on conversational user interfaces
- IUI – conference on intelligent user interfaces
- HAI – conference on human agent interaction
- HRI – conference on human robot interaction
- INTERSPEECH – conference on spoken language processing
- IVA – conference on intelligent virtual agents
- LREC – conference on language resources and evaluation
- SemDial – workshop on the semantics and pragmatics of dialogue
- SIGdial – special interest group on discourse and dialogue
- Dagstuhl Seminar 20021 – SLIVAR – spoken language interaction with virtual agents and robots
- Dagstuhl Seminar 21381 – Trust-CA – conversational agent as trustworthy autonomous system

The need for strengthened collaboration between academic research and industry. There seems to be a disconnect between large commercial vendors and the academic community regarding research of relevance to the technological basis for CAs. Furthermore, a shift may be observed where high profile research increasingly is coming from large technology companies. Along with this, data and computational resources are increasingly isolated within commercial entities.

In consequence of "data as the new code", there is a need for researchers to access data to fully understand or replicate a system or research conducted on a system. However, challenges exist for sharing of data held by industry, including privacy risks and difficulties in sanitizing data at scale. Also, there may be a perceived loss of competitive advantage in sharing data resources.

Furthermore, there seems to be a talent-pipeline challenge in the AI space in which it is difficult for academia to attract and keep PhDs and postdocs due to the imbalance in financial compensation between these positions and the roles available within industry.

The challenge for industrial players to oversee and evaluate CAs. Automatic and comprehensive evaluation of CAs is technically challenging. There is a need for better evaluation methods for CA owners. The availability of tools, frameworks, and platforms has reduced barriers for uptake of CAs in industry. At the same time, there may be a lack of sufficient guidelines for practitioners within industry using these tools, e.g., for intent design and optimization. Hence, while creating a CA may be quick and low-effort, it can be challenging to design and develop a CA of sufficient quality to provide the desired user experience.

Possibly, CA owners relying on tools with insufficient documentation, guidelines, and transparency may be unaware of the limitations of component technologies and thus experience overtrust in those tools.

Investigation is required to establish best practice guidelines in this space. Specific guidelines will vary from one platform to another due to differences in model architecture, training data, and other platform features and modules such as entity recognition or sentiment detection.

Due to the black-box character and evolving nature of platform components, this may be something that needs to be done by platform owners. Furthermore, there is a need for confidence scores for component technologies to allow CA owners to build trust in their systems.

System architecture and complexity. An aspect of CA systems that makes it challenging to manage trust is system architecture and complexity. This complexity concerns, in part, end-to-end systems and large language models. As, for example, seen in challenges of handling bias in data and system output, which is important to a trustworthy system. Due to the inherent complexity in such systems, managing trust may come at a cost (e.g. accuracy). Complex modular systems need to spread trust along the chain of modules but tuning one component might affect another. Possibly, certifications may be developed to handle seemingly competing objectives in complex CA systems.

Ethics and transparency. The AI and technical development underpinning of CAs also entail a range of ethical issues. Tools and approaches such as emotion detection can have great benefit and be used in personalization but while some use cases can be ethical, other scenarios may be ethically questionable. One approach to addressing such ethical issues may be to look towards other fields that have faced similar challenges.

Transparency can increase trust in CAs. A CA's behaviour should be transparent. That is, it needs to be understandable but also to allow for in-depth insights, e.g. into the used data sources. The need for transparency may, however, vary between user groups. Hence, to achieve transparency in CAs user's roles and profiles need to be considered. CAs may also need to afford transparency with different modalities of interaction.

Conversational repair and trust. Conversational repair is important in CAs, to support needed adjustment or adaptation of dialogue to mitigate interpretational issues or misunderstandings. Repair strategies impact user trust and attitudes towards a chatbot. Detecting the need for conversational repair may be challenging and we currently lack sufficient automated approaches – for example in the case of false positive replies in CAs.

4.5.3 Future Research

Relevant next steps in research towards strengthening the technological basis for trust include:

- Strengthen opportunities for collaboration between academic research and industry, including how to share data or metadata when publishing technical research on CAs
- Develop guidelines for design, development, and evaluation of CAs, for use of all human actors in the CA supply chain.
- Research efforts addressing how to manage trust in complex systems enabling current and future CAs
- Research towards transparency and explainability in CAs
- Research addressing automatic conversational repair in CAs

4.6 Breakout Group 5: Definition, Conceptualization and Measurement of Trust

Group 5

Contributors:
Group 5a (in person): Martin Porcheron, Minha Lee, Birthe Nesset, Frode Guribye
Group 5b (online): Margot van der Goot, Roger K. Moore, Ricardo Usbeck, Ana Paiva, Catherine Pelachaud, Elayne Ruane
Group 5c (in person): Björn Schuller, Guy Laban, Dimosthenis Kontogiorgos, Matthias Kraus, Asbjørn Følstad

4.6.1 Goal and key questions

Goal: Enable assessment and measurement of trust in CAs
Key question: How to define, conceptualize and measure trust in CA?
Relevant aspects: Modelling frameworks – antecedents / consequents; basis in knowledge on human-human communication; psychology of trust – over-trust / intuition

4.6.2 Key insights

Defining trust. Trust is addressed in different disciplines, both as a general concept within psychology, sociology, and management research (e.g., Rousseau, 1998; Mayer et al., 1995) and – more recently – as a term of relevance for users' perceptions of technology (e.g., Corritore et al., 2003; McKnight et al., 2011). A range of definitions exists for trust. There is variation in definitions concerning whether trust should be construed as a belief or attitude (Lewis et al., 2018; Lee & See 2004), and the degree to which there is a behavioural element in trust (Söllner et al., 2016; Malle & Ullman, 2021).

For conceptual clarity, it may be useful to consider trust an attitude which may be founded by trusting beliefs, and which may lead to trusting behaviour.

Trusting behaviour is determined by trust and may as such be an indicator of trust – provided users have a choice. Trusting behaviour is also moderated by environment, user group, and use case. An example of trusting behaviour is self-disclosure. Trust may also impact engagement level in behaviour and tendency to repeated use.

Developing trust through conversational interactions. The notion of trust in technology arguably is of particular relevance to CAs, due to their conversational interaction with users. Conversations are humanlike which has implications for users trusting beliefs and behaviours. Furthermore, conversations may be relational, leading to expectations of evolving capabilities in agent. Conversations may also be cooperative, leading to expectations of mutual adaptations in the user and conversational agent to achieve a common goal.

On this background, trust in CA may be considered as gradually built through conversations. In consequence, four trust concepts may be of particular relevance for CAs:

- *Initial trust*: trust required for users to initiate interaction. Initial trust corresponds to the notion of calculating trust in Rousseau et al. (1998)

- *Sharing trust*: trust required for sharing information with chatbot. The relevance of sharing trust may depend on varying levels of perceived sensitivity in the domain or topic of CA interaction.
- *Reliance trust*: trust required for relying on chatbot recommendations or decision support, that is, trust impacting user beliefs or behaviour beyond the context of the CA interaction.
- *Long-term trust*: trust required for repeated / routine use. Long term trust corresponds to the notion of relational trust in Rousseau et al. (1998)

Extending the trust model of Rousseau et al. (1998), the four trust concepts for CAs may be mapped out on a timeline of the evolving relation between user and CA as follows:

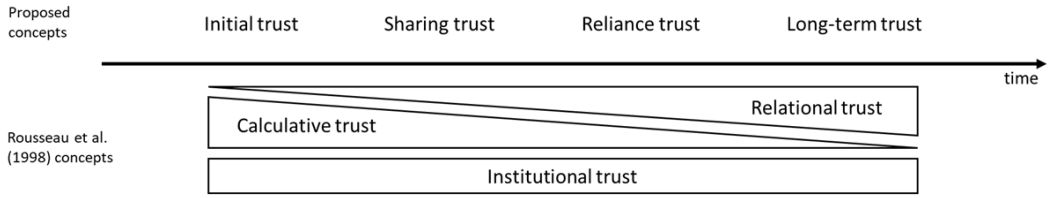

Figure 4 Extended trust model.

Balancing trust and trustworthiness. When considering trust, it is critical to distinguish between perceived trust and trustworthiness.

Perceived trust is held by the trustor, typically the user. Perceived trust and related trust beliefs may be measured through a range of self-report measurements, for example from information systems research (e.g. Lankton et al., 2015), social robotics (c.f. review in Hancock et al., 2020). Perceived trust may be impacted by the trustworthiness of the trustor. However, as information on this may not be available, other characteristics may impact trust. For CAs, anthropomorphism may be such a characteristic, as it may impact trust though not be correlated with trustworthiness.

Trustworthiness is a characteristic of the trustee, typically the service provider. Trustworthiness may depend on factors such as transparency, reliability, consistency, sincerity, honesty, integrity, benevolence, competence, and cooperation. These factors, though not necessarily static, may be considered observable characteristics in a trustee.

There is a need to study trustworthiness and perceived trust in parallel – to address potential overtrust (low trustworthiness and high perceived trust) and undertrust (high trustworthiness and low perceived trust). There is a lack of approaches or measurements for the integrated study of trustworthiness and trust.

Measuring trust by integrating self report measures and behavioural measures. In existing scales and measurements, trust is typically construed as personal, mainly available to researchers through self report measurements. Nevertheless, trust can be interpreted as reflected in and through people's behaviour, rather than merely a stance prior to the use of some device or system. Trust as reflected in trusting behaviour may enables trust to be measured also on the basis of user behaviour. There seem however to be a lack of distinct behaviour scales for trust assessment.

Possibly, trust may be measured by having a CA asking about sensitive information and monitor users' disclosing behaviour. Specifically, a tiered approach may be useful, based on asking questions of personal information of increasing level of sensitivity to infer a person's

level or trust. However, the choice of behavioural measures of trust may depend on the context of the CA.

An integrated approach, combining self-report measures and tiered behavioural measures seems a promising approach for future research.

A proposed integrating framework for measuring trust and trusting behaviour. Following from the above, instruments and data sources for measuring trust may be divided into two broad groups: Subjective and objective measures:

- *Subjective measures* concern the measurement of trust determinants / trusting beliefs or behavioural intent (e.g., Lankton et al., 2015; Yagoda & Gillan, 2012). As a subjective measurement, perceptions of trust are expected to be explicit from the subject's report, corresponding to the subject's trust beliefs. Nevertheless, these might not be consistent with the subject's trusting behaviour due to personal perceptions and attitudes of the subject regarding the conversational AI system – e.g., due to scepticism of AI (Araujo et al., 2020).
- *Objective measures* include measures of physiological states, speech / voice, interaction with agent (e.g., sharing behaviour), changes in beliefs due to agent, behaviour in the world due to agent. Accordingly, the subject's behaviour would implicitly indicate higher or lower levels of trust. THe association between trusting behaviour and trust should be studied individually, depending on context, settings and task. Within the scope of conversational AI, behaviour such as self-disclosure (e.g., Laban et al., 2021a), reciprocity (e.g., Zonca et al., 2021), and changes in disclosure and expression over time (e.g., Laban et al., 2021b) could implicitly indicate changes in trust. These behaviours, however, might not be consistent with one's trust beliefs due to, for example, habits and needs (e.g., having the need to share, or being an impulsive individual) or affect-based factors of trust like the system's heuristics and demonstrated social cues (e.g., one might be more likely to share information with a more persuasive system despite not trusting it; e.g., Ghazali et al., 2019).

Subjective and objective measures may be included in a framework of trusting beliefs and trusting behaviour as follows:

4.6.3 Future Research

The following questions require further research effort to address:

- Developing a comprehensive framework to capture how trust evolves across long-term use.
- Refining the framework for trusting beliefs and trusting behaviour.
- Developing integrated approaches and measures for studying users perceived trust and the trustworthiness of service providers, to mitigate overtrust and undertrust.
- Developing integrated measures of trust and trusting behaviour, combining self report measures and tiered behavioural measures to support standardised measure for trust in conversational agents, and incorporating this in conversational systems.

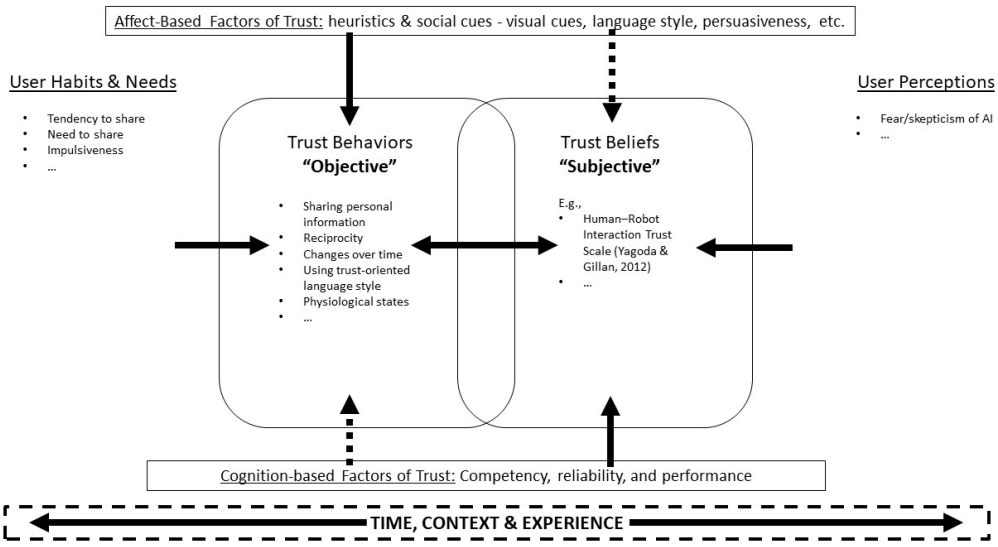

Figure 5 Framework of trusting beliefs and trusting behaviour.

References

1 Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI and Society, 35(3), 611 – 623.

2 Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. International Journal of Human-Computer Studies, 58(6), 737-758.

3 Ghazali, A. S., Ham, J., Barakova, E., & Markopoulos, P. (2019). Assessing the effect of persuasive robots interactive social cues on users' psychological reactance, liking, trusting beliefs and compliance. Advanced Robotics, 33(7 – 8), 325 – 337.

4 Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2020). Evolving trust in robots: specification through sequential and comparative meta-analyses. Human factors, 0018720820922080.

5 Laban, G., George, J.-N., Morrison, V., & Cross, E. S. (2021a). Tell me more! Assessing interactions with social robots from speech. Paladyn, Journal of Behavioral Robotics, 12(1), 136 – 159.

6 Laban, G., Kappas, A., Morrison, V., & Cross, E. S. (2021b). Protocol for a Mediated Long-Term Experiment with a Social Robot. PsyArXiv.

7 Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. Journal of the Association for Information Systems, 16(10).

8 Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

9 Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In Foundations of trusted autonomy (pp. 135-159). Springer, Cham.

10 Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In Trust in Human-Robot Interaction (pp. 3-25). Academic Press.

11 Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of management review, 20(3), 709-734.

12 McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. ACM Transactions on management information systems (TMIS), 2(2), 1-25.
13 Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. Academy of Management Review, 23(3), 393-404.
14 Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. European Journal of Information Systems, 25(3), 274-287.
15 Yagoda, R. E., & Gillan, D. J. (2012). You Want Me to Trust a ROBOT? The Development of a Human – Robot Interaction Trust Scale. International Journal of Social Robotics, 4(3), 235 – 248.
16 Zonca, J., Folsø, A., Sciutti, A. (2021). The role of reciprocity in human-robot social influence. iScience, 24(12), 103424.

4.7 Breakout Group 6: Interaction Design

Group 6

Contributors:
Group 6a (in person): Oliver Bendel, Sebastian Hobert, Ryan Schuetzler, Elisabeth André, Leigh Clark, Clayton Lewis, Stefan Schaffer, Eren Yildiz, Effie Law
Group 6b (online): Stefan Morana, Heloisa Candello, Christine Liebrecht, Zhou Yu, Dakuo Wang, Michelle Zhou, Ana Paula Chaves, Cosmin Munteanu, Soomin Kim

4.7.1 Goal and key questions

Goal: Identify interaction designs to strengthen trust in CA.
Key questions: How to design trusted conversational user interfaces?
Relevant aspects: UX – human-AI sociability; Human-in-the-loop; Evaluation – reliability/acceptance; Group interaction.

4.7.2 Key insights

4.7.2.1 Group 6a

The group started with reflecting on the following aspects of trust:

- Brand and UX: The producer of a chatbot affects our perception of trust; we trust certain products because we trust certain brands; the implication of UX-trust relation
- Group effect: If we trust people, and those people trust a chatbot, then we are more likely to trust it as well.
- Domain-dependency: Certain domains are more sensitive to trust fluctuation
- Modality: Intricate relations between modality, risk and trust

Next, the group focused on some specific aspects of conversational interactions. The multifaceted nature of trust and the numerous factors of people's interactions with CAs that can impact on perceptions and behaviours, complicating our understanding of how, why and when to design for trust and subsequently evaluate it. We present a discussion of some critical aspects of CA interactions and highlight the need for a holistic approach to creating trustworthy CAs.

- **Multimodality** A critical question is whether multimodality can increase or decrease trust. On the other hand, multimodality might help to increase accuracy (i.e. automatic emotion recognition) which might help to increase trust in CAs. On the other hand, multimodality could decrease trust (privacy issues due to permission requests to webcam or other sensors). It depends on whether requesting permissions to access multiple sensors, e.g., on a phone, might test the trust of users, especially in the case of iPhone where each sensor is requested in sequence might annoy users. Nonetheless, it could creep-out users if the multimodal information is used in inappropriate or clumsy ways, especially when the verbal information conflicts with the nonverbal (e.g., "You say that you are happy, but your voice sounds sad. Are you depressed?"). Here below we discuss several aspects of multimodality:
  - **Preferences**: Cultural aspects can lead to different preferences among users. Some users may not like to use voice but text input only (if this is possible). An adaptation to such individual preferences should be considered during CA design. According to individual differences of users, a customization would be desirable as some people might prefer different modalities for interaction. Preferences may include choosing the voice, the tone of the voice, or the formal/informal style.
  - **Input**: An important question for future research is how multimodal sensory perception can be used to enhance CA effectiveness/accuracy. Depending on the use case, different multimodal sensors could be used to improve the interaction, including keyboard, camera, microphone, as well as accelerometer, thermal sensor, GSR and others. A fusion of the information coming from modality specific modules should generate a more reliable intention detection.
  - **Output generation**: The answers from the system have to be output using the appropriate modality. Usually a symmetry between the input and the output modality is expected by the user. The output generation module has to prepare the system feedback for the necessary modalities. This can include text generation, speech synthesis or graphic generation. When generating output text, the CA often has to integrate database answers into output text. Thereby errors can occur while producing the correct form of the word(s), e.g. if it's singular or plural, and the correct cases (Genitive, Dative or Accusative), or verb form. Today mostly templates are used to generate output prompts. Neural methods taking into account the integration of such database results are not yet mature.
- **Transparency** Trust might be fostered if the CA provides explanations about what it is capable of doing or understanding. The relation between trust and transparency: Is it reasonable to assume the more transparency we have, the more trust we get? Feedback from the chatbot should be personalized. If I want shorter feedback, the chatbot should do it so. Furthermore, other aspects of multimodality have to be considered:
  - Explaining why certain permissions might be needed: Do we trust the explanation?
  - Do explanations matter? Are too many permissions/explanations detrimental to trust or acceptance?
  - Baseline level based on your general preferences
- **Voice and Language** There are numerous features of CA speech that can impact on people's perceptions and behaviours. Features of voice quality, "those characteristics which are present more or less all the time that a person is talking' (Abercrombie, 1967, p. 91 in Laver, 1980, p. 1)," include an agent's perceived accent, gender, age, prosody and human-likeness. In addition to voice quality, the linguistic content delivered by a

CA can have similar impacts. Examples include language, dialect, register and style (e.g. Bendel 2018).

- **Context** Using context information (e.g., in learning contexts) might help to provide more accurate answers. If context information is missing, it might be annoying for users.
  - Application context (e.g., health, mental health, customer service)
  - Environmental context (e.g., room, building)
  - Social context
- **Evaluation**
  - A user-centered design process is important. Co-design or participatory design or human-centred approaches will help.
  - Questionnaire including trust related scales: e.g., https://ueqplus.ueq-research.org
  - How to measure trust using questionnaires and without questionnaires? Is it possible?
- **Should a CA Be Humanlike?** What is humanlikeness? Is it the ability of the bot to sound human, talk like a human? Or the ability to do what a human would do? The humanlikeness of the CA, at least insofar as it does not enter the uncanny valley, is likely to increase trust as long as the bot is upfront about its botness. Alexa's voice and capabilities could improve to the point of being completely humanlike. This may be related to the notion of partner models, which "refer...to a person's internal representation of an interlocutor's (human or machine) dialogic competence" (Doyle et al., 2021). Some studies have pointed out situations in which a more humanlike agent underperforms compared to a less humanlike agent with respect to a desired outcome (e.g., Schuetzler et al. 2018). These findings at least suggest that humanlikeness and its consequences are not universally desirable.

4.7.2.2 Group 6b

The main points of the discussion on the key question "How to design trusted conversational user interface" are categorised summarised in the following:

- *Domain*: design of CA is domain-dependent, as shown in examples: tourism, education for early childhood, financial, healthcare, informal public spaces such as museum
- *Transparency*: Explainable AI; Personal identifiable data storage (what do you know about me); Split the content in small chunks/topics; Strategies to show many options – personalised recommendations tailored by interest
- *Chatbot language design*: Humanlike design increases frustration; Register theory (age, location, language style); Infrastructure behind the chatbot
- *Accuracy*: answer as expected
- *Relationship*: engagement, satisfaction
- *Voice and text interfaces*: Speech interfaces can have higher cognitive load than text ones, depending on the task; Text interface – privacy information
- *Conversation flow*: Proactive vs. reactive bot; Decision-making system vs. informational bots; Disambiguation; Repair strategies; Multi-bot vs. single-bot
- *Settings*: privacy and public settings
- *Development*: technical devices usability and bugs

4.7.3 Future Research

The following research topics on CA interaction design can be explored as future work:

- Individual differences – configurable preferences are one way to adapt an agent to individual differences, but we must remember that trends/correlations/construct relationships are typically studied in aggregate, but individuals vary significantly from the mean.
- Identify which research findings that are generalizable across a variety of contexts and which are limited to within some specific context
- Potential limitations and ethical considerations of imitating human-likeness in CA design
- Resolving conflicts from multimodal sensors
- Impact of different styles/levels of embodiment (e.g. robotics, virtual avatars) on trust
- How best to appropriately evaluate the impact of interaction design choices on trust

References

1 Abercrombie, D. (1967). Elements of general phonetics (Vol. 203). Edinburgh: Edinburgh University Press.
2 Bendel, Oliver. (2018). From GOODBOT to BESTBOT. In: The 2018 AAAI Spring Symposium Series. AAAI Press, Palo Alto 2018. pp. 2-9.
3 Doyle, P. R., Clark, L., & Cowan, B. R. (2021, May). What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14).
4 Laver, J. (1980). The phonetic description of voice quality: Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.
5 Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F. (2018). The Influence of Conversational Agent Embodiment and Conversational Relevance on Socially Desirable Responding. Decision Support Systems, 114(October), 94 – 102.

5 Open problems

5.1 Trust-CA: Conclusion and Suggested Readings

Trust-CA All

5.1.1 Conclusion

Through the twenty talks, six breakout groups and informal discussions, the Seminar's attendees explored the topic of Trust-CA widely as well as deeply. As the field is emerging, there are still many questions to answer, as shown in the report of each of the breakout groups. Among them, ethics of CA is a key concern. In fact, in the pre-Seminar survey, many of the respondents mentioned different aspects of ethics pertaining to CA and other AI-infused autonomous systems. Ethical considerations are highly relevant to the three main challenges for the Seminar (see Executive Summary). While the outcomes of the Seminar can provide insights to resolving these challenges, more research efforts are required. Encouraging dialogues and collaborations among different research communities working on conversational agents is essential for advancing this field. The Seminar Trust-CA has made a critical step along this direction.

5.1.2 Suggested Readings

In moving forward, it is necessary to review what has achieved in the past through reading the related publications. Prior to the seminar, the organizers asked the attendees to list their recommended readings of relevance to trust in conversational agents. The following readings were suggested.

References

1 Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & Society.
2 Batliner, A., Hantke, S., & Schuller, B. W. (2020). Ethics and good practice in computational paralinguistics. IEEE Transactions on Affective Computing.
3 Bendel, O. (2018) From GOODBOT to BESTBOT. In The 2018 AAAI Spring Symposium Series (pp. 2-9). Palo Alto, CA: AAAI Press.
4 Bendel, O. (2020). The morality menu project. In M. Nørskov, J. Seibt, O. S. Quick, (Eds.). Culturally Sustainable Social Robotics – Challenges, Methods and Solutions: Proceedings of Robophilosophy 2020. (pp. 257-268). Amsterdam, Netherlands: IOS Press
5 Bendel, O., Schwegler, K., Richards, B., (2017). Towards Kant machines. In The 2017 AAAI Spring Symposium Series (pp. 7-11). Palo Alto, CA: AAAI Press.
6 Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. IEEE Transactions on Human-Machine Systems, 50(3), 215 – 224.
7 Biancardi, B., Dermouche, S., & Pelachaud, C. (2021). Adaptation mechanisms in human-agent interaction: Effects on User's Impressions and Engagement. Frontiers in Computer Science.
8 Bodó, B. (2020). Mediated trust: A theoretical framework to address the trustworthiness of technological trust mediators. New Media & Society.
9 Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., and Cowan, B. R. (2019). What makes a good conversation? challenges in designing truly conversational agents. In Proceedings of CHI'19. New York, NY: ACM.
10 Cranshaw, J., Elwany, E., Newman, T., Kocielnik, R., Yu, B., Soni, S., ... & Monroy-Hernández, A. (2017). Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In Proceedings of CHI'17. (pp. 2382-2393). New York, NY, ACM.
11 Cross, E. S., & Ramsey, R. (2021). Mind Meets Machine: Towards a Cognitive Science of Human-Machine Interactions. Trends in Cognitive Sciences, 25(3), 200 – 212.
12 De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. Journal of Experimental Psychology: Applied, 22(3), 331 – 349.
13 De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human – machine interaction. Ergonomics, 61(10), 1409 – 1427.
14 Devillers, L., Kawahara, T., Moore, R. K., & Scheutz, M. (Eds.) (2020). Spoken language interaction with virtual agents and robots (SLIVAR): Towards effective and ethical interaction (Dagstuhl Seminar 20021). Dagstuhl Reports (Vol. 10). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum fur Informatik.
15 Falcone, R., & Castelfranchi, C. (2004). A belief-based model of trust. In Trust in Knowledge Management and Systems in Organizations (pp. 306-343). IGI Global.
16 Fan, X., Chao, D., Zhang, Z., Wang, D., Li, X., Tian, F. (2021) Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case study. Journal of Medical Internet Research, 23(1).

17 Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. International Journal of Human-Computer Studies, 132, 138 – 161.

18 Følstad A., Nordheim C.B., & Bjørkli C.A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In Proceedings of the International Conference on Internet Science – INSCI 2018. LNCS, vol 11193. Cham, Switzerland: Springer.

19 Følstad, A., Araujo, T., Papadopoulos, S., Law, E. L. C., Luger, E., Goodwin, M., & Brandtzaeg, P. B. (Eds.). (2021). Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23-24, 2020, Revised Selected Papers (Vol. 12604). Springer Nature.

20 Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. In International Symposium on Collaborative Technologies and Systems (pp. 106-114)

21 Fussell, S. R., Kiesler, S., Setlock, L. D., & Yew, V. (2008). How people anthropomorphize robots. In 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 145-152). New York, NY: ACM.

22 Garfinkel, H. (1963). A conception of and experiments with "trust" as a condition of stable concerted actions. In O. J. Harvey (Ed.), Motivation and Social Interaction: Cognitive Determinants (pp. 187-238). Ronald Press.

23 Gefen, D., & Straub, D. W. (2004). Consumer trust in B2C e-Commerce and the importance of social presence: experiments in e-Products and e-Services. Omega, 32(6), 407 – 424.

24 Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In Proceedings of CHI'19. New York, NY: ACM.

25 Han, X., Zhou, M. X., Turner, M., & Yeh, T. (2021). Designing effective interview chatbots: Automatic chatbot profiling and design suggestion generation for chatbot debugging. In Proceedings of CHI'21. New York, NY: ACM.

26 Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., & Wong, B. L. (2019). Algorithmic transparency of conversational agents. In Joint Proceedings of the ACM IUI 2019 Workshops. CEUR Workshop Proceedings.

27 Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407 – 434.

28 Jaisie S., Munteanu, C., Ramanand, N., & Tan, Y. R. (2021). VUI influencers: How the media portrays voice user interfaces for older adults. In CUI 2021 – 3rd Conference on Conversational User Interfaces (Article 8), New York, NY: ACM,

29 Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. In Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2). New York, NY: ACM.

30 Komiak, S. Y., & Benbasat, I. (2006). The effects of personalization and familiarity on trust and adoption of recommendation agents. MIS Quarterly, 941 – 960.

31 Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. In Proceedings of Mensch und Computer 2019 (MuC'19) (pp. 31-42). New York, NY: ACM.

32 Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, Humanness, and Trust: Rethinking Trust in Technology. Journal of the Association for Information Systems, 16(10), 880 – 918.

33 Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

34 Lee, Y. C., Yamashita, N., & Huang, Y. (2020). Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. In Proceedings of the ACM on Human-Computer Interaction, 4(CSCW1). New York, NY: ACM.

35 Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In Foundations of trusted autonomy (pp. 135-159). Cham, Switzerland: Springer.

36 Lucas, G., Stratou, G., Lieblich, S., & Gratch, J. (2016). Trust me: multimodal signals of trustworthiness. In ICMI '16: Proceedings of the 18th ACM International Conference on Multimodal Interaction (pp. 5-12). New York, NY: ACM.

37 Marche, S. (2021). The Chatbot Problem. The New Yorker, July 23, 2021.

38 Marge, M., Espy-Wilson, C., Ward, N. G., Alwan, A., Artzi, Y., Bansal, M., ... & Yu, Z. (2021). Spoken language interaction with robots: Recommendations for future research. Computer Speech & Language.

39 Meng, L. & Schaffer, S. (2020). A reporting assistant for railway security staff. In Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20) (article 31). New York, NY: ACM.

40 Moon, Y. (2000). Intimate exchanges: Using computers to elicit self-disclosure from consumers. Journal of Consumer Research, 26(4), 323 – 339.

41 Moore, R. K. (2017). Appropriate voices for artefacts: some key insights. In 1st Int. Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017) (pp. 7-11). Skovde, Sweden: VIHAR.

42 Moradinezhad, R., Solovey, E.T. (2021). Investigating trust in interaction with inconsistent embodied virtual agents. International Journal of Social Robotics.

43 Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies, 27(5-6), 527 – 539.

44 Munteanu, C.; Benett, A.; Rafih, H., Liaqat, A., & Aly, Y. (2018) Designing for older adults: Overcoming barriers to a supportive, safe, and healthy retirement. Wharton Pension Research Council Working Papers.

45 Nickel, Philip J. (2013). Trust in technological systems. In: M. J. de Vries, S. O. Hansson, & A. W. Meijers (eds.): Norms in Technology (pp. 223-237). Dordrecht, Netherlands: Springer

46 Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230 – 253.

47 Parviainen, E., & Søndergaard, M. L. J. (2020). Experiential qualities of whispering with voice assistants. In Proceedings of CHI'20. New York, NY: ACM.

48 Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018, April). Voice interfaces in everyday life. In Proceedings of CHI'18 (paper 640). New York, NY: ACM.

49 Rheu, M., Shin, J. Y., Peng, W., & Huh-Yoo, J. (2021). Systematic review: trust-building factors and implications for conversational agent design. International Journal of Human-Computer Interaction, 37(1), 81 – 96.

50 Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the trust perception scale-HRI. In Robust Intelligence and Trust in Autonomous Systems (pp. 191-218). Springer, Boston, MA.

51 Schaefer, K. E., Chen, J. Y., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors, 58(3), 377 – 400.

52 Shneiderman, B. (2020). Human-centered artificial intelligence: Trusted, reliable & safe. International Journal of Human – Computer Interaction, 36(6), 495 – 504.

53 Sin, J., & Munteanu, C. (2020). An empirically grounded sociotechnical perspective on designing virtual agents for older adults. Human – Computer Interaction, 35(5-6), 481 – 510.

54 Skjuve, M., & Brandtzæg, P. B. (2018). Chatbots as a new user interface for providing health information to young people. In Youth and News in a Digital Media Environment – Nordic-Baltic Perspectives (pp. 59-66). Gothenburg, Sweden: Nordicom.

55 Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In Proceedings of the Technology, Mind, and Society. New York, NY: ACM.

56 Van Pinxteren, M. M. E., Pluymaekers, M., & Lemmink, J. G. A. M. (2020). Human-like communication in conversational agents: a literature review and research agenda. Journal of Service Management, 31(2), 203 – 235.

57 Wang, L., Wang, D., Tian, F., Peng, Z., Fan, X., Zhang, Z., Yu, M., Ma, X., & Wang, H. (2021) CASS: Towards building a social-support chatbot for online health community. In Proceedings of the ACM on Human-Computer Interaction 5(CSCW1) (article 9). New York, NY: ACM.

58 Xiao, Z., Zhou, M. X., Chen, W., Yang, H. & Chi, C. (2020). If I hear you correctly: Building and evaluating interview chatbots with active listening skills. In Proceedings of CHI'20. New York, NY: ACM.

59 Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. Computers & Education, 161.

60 Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents: The effect of personality. ACM Transactions on Interactive Intelligent Systems (TiiS), 9(2-3).

Participants

- Elisabeth André
Universität Augsburg, DE
- Oliver Bendel
FH Nordwestschweiz – Windisch, CH
- Leigh Clark
Swansea University, GB
- Asbjørn Følstad
SINTEF – Oslo, NO
- Frode Guribye
University of Bergen, NO
- Sebastian Hobert
Georg August Universität – Göttingen, DE
- Andreas Kilian
Universität des Saarlandes – Saarbrücken, DE
- Dimosthenis Kontogiorgos
KTH Royal Institute of Technology – Stockholm, SE
- Matthias Kraus
Universität Ulm, DE
- Guy Laban
University of Glasgow, GB
- Effie Lai-Chong Law
Durham University, GB
- Minha Lee
TU Eindhoven, NL
- Clayton Lewis
University of Colorado – Boulder, US
- Birthe Nesset
Heriot-Watt University – Edinburgh, GB
- Catherine Pelachaud
Sorbonne University – Paris, FR
- Martin Porcheron
Swansea University, GB
- Stefan Schaffer
DFKI – Berlin, DE
- Ryan Schuetzler
Brigham Young University – Provo, US
- Björn Schuller
Universität Augsburg, DE
- Eren Yildiz
University of Umeå, SE

Remote Participants

- Theo Araujo
University of Amsterdam, NL
- Susan Brennan
Stony Brook University, US
- Heloisa Candello
IBM Research – Sao Paulo, BR
- Ana Paula Chaves
Federal University of Technology – Paraná, BR
- Cristina Conati
University of British Columbia – Vancouver, CA
- Benjamin Cowan
University College – Dublin, IE
- Laurence Devillers
CNRS – Orsay, FR & Sorbonne University – Paris, FR
- Jasper Feine
KIT – Karlsruher Institut für Technologie, DE
- Jonathan Grudin
Microsoft – Redmond, US

- Evelien Heyselaar
Radboud University
Nijmegen, NL
- Soomin Kim
Seoul National University, KR
- Stefan Kopp
Universität Bielefeld, DE
- Yi-Chieh Lee
NTT – Kyoto, JP
- Oliver Lemon
Heriot-Watt University –
Edinburgh, GB
- Q. Vera Liao
IBM TJ Watson Research Center
– White Plains, US
- Christine Liebrecht
Tilburg University, DE
- Roger K. Moore
University of Sheffield, GB
- Stefan Morana
Universität des Saarlandes –
Saarbrücken, DE
- Cosmin Munteanu
University of Toronto
Mississauga, CA
- Ana Paiva
INESC-ID – Porto Salvo, PT
- Symeon Papadopoulos
CERTH – Thessaloniki, GR
- Caroline Peters
Universität des Saarlandes –
Saarbrücken, DE
- Rolf Pfister
Cognostics – Pullach, DE
- Olivier Pietquin
Google – Paris, FR
- Aleksandra Przegalinska
Kozminski University, PL
- Elayne Ruane
University College Dublin, IE
- Marita Skjuve
SINTEF – Oslo, NO
- Cameron Taylor
Google – London, GB
- Ricardo Usbeck
Universität Hamburg, DE
- Margot van der Goot
University of Amsterdam, NL
- Dakuo Wang
IBM T.J. Watson Research
Center – Yorktown Heights, US
- Saskia Wita
Universität des Saarlandes –
Saarbrücken, DE
- Levi Witbaard
OBI4wan – Zaandam, NL
- Zhou Yu
Columbia University –
New York, US
- Michelle X. Zhou
Juji Inc. – Saratoga, US