What Do WEIRD and Non-WEIRD Conversational Agent Users Want and Perceive? Towards Transparent, Trustworthy, Democratized Agents

JESSICA VAN BRUMMELEN, Massachusetts Institute of Technology, USA
MAURA KELLEHER, Massachusetts Institute of Technology, USA
MINGYAN CLAIRE TIAN, Wellesley College, USA
NGHI HOANG NGUYEN, Massachusetts Institute of Technology, USA

Fig. 1. A portion of children from non-WEIRD countries’ responses during an ideation design session about their ideal conversational agents. Other responses came from parents from non-WEIRD countries, and parents and children from WEIRD countries.

A majority of researchers who develop design guidelines have WEIRD, adult perspectives. This means we may not have technology developed appropriately for people from non-WEIRD countries and children. We present five design recommendations to empower designers to consider diverse users’ desires and perceptions of agents. For one, designers should consider the degree of task-orientation of agents appropriate to end-users’ cultural perspectives. For another, designers should consider how competence, predictability, and integrity in agent-persona affects end-users’ trust of agents. We developed recommendations following our study, which analyzed children and parents from WEIRD and non-WEIRD countries’ perspectives on agents as they create them. We found different subsets of participants’ perceptions differed. For instance, non-WEIRD and child perspectives emphasized agent artificiality, whereas WEIRD and parent perspectives emphasized human-likeness. Children also consistently felt agents were warmer and more human-like than parents did. Finally, participants generally trusted technology, including agents, more than people.

CCS Concepts: • Human-centered computing → Natural language interfaces; User models; User interface programming; • Social and professional topics → Children; Age; Cultural characteristics; K-12 education; • Computing methodologies → Intelligent agents.

Additional Key Words and Phrases: conversational agents, chatbots, virtual assistants, conversational AI, non-WEIRD and WEIRD, parents, trust, partner models, agent personification, computational action, technology democratization
1 INTRODUCTION

Conversational artificial intelligence (AI)—or the ability of a computer program to understand human language and respond accordingly—is ripe with potential. Imagine a conversational agent engaging children in learning history with a virtual Rosa Parks; an agent writing a persuasive article for you after you have a natural-language discussion; an agent providing constant, accurate healthcare answers to those in need. With recent major advances in natural language processing and automatic speech recognition, including transformers, transfer learning and foundation models, these ideas are not far-fetched [6, 8, 13, 19, 23, 28, 58]. It is not difficult to envision natural language becoming the primary way we communicate with computers.

Nonetheless, current agents, like Google Home, Apple’s Siri and Amazon Alexa, still misrecognize speech and misunderstand intent [4, 4, 51]. For instance, researchers found speech recognition systems by Amazon, Google, IBM and Microsoft did substantially worse when recognizing black speakers versus white [27]. Others have found significant gender biases in word embeddings, causing words like “homemaker” to be heavily associated with “woman”, and “computer programmer” with “man” [5, 64]. Biases in AI systems are widespread, being found in medical technologies, justice systems, advertising, policing systems, and elsewhere [45]. If users are not aware of such flaws, there could be serious implications, including misinformation being spread, human bias being compounded, and users unwittingly acting on incorrect advice.

Ideally, agents would be developed to portray the reality of their abilities and limitations to their users through effective design. In a study with AI decision-aids, researchers describe how if users are too averse to technology’s advice and information, they cannot truly benefit from using the technology. However, if they are too appreciative, users may make ill-informed decisions when technology presents incorrect information [18]. By portraying conversational agents in an honest way through design, discrepancies between users’ expectations of agents—or their agent “partner models”—and the reality of agents can be reduced, which can also reduce user frustration [15].

In our study, we investigate users’ partner models of agents, as well as various aspects of their trust of agents. The results revealed how—for many aspects of agent partner models and trust—users of different backgrounds (e.g., locations, prior experiences, ages, etc.) perceived agents differently. Participants’ general trust of agents’ correctness (compared to other people’s and systems’ correctness), however, was similar for all major subsets of participants. In general, people trusted agents more than their friends and parents. Based on these results, we developed agent design recommendations to foster appropriate levels of trust of agents, and accurate partner models.

Historically, human-computer interaction research has largely recruited participants from Western, Educated, Industrialized, Rich and Democratic (WEIRD) countries, who comprise less than 12% of the world’s population [20, 29]. This means many of the design recommendations developers use are likely biased towards this population. Furthermore, a large portion of software developers reside in WEIRD countries [16, 25, 60], meaning technology development is likely further biased towards the WEIRD population. In order to address this, and develop technology meaningful and relevant to more of the world, researchers have developed different strategies. One strategy involves including more participants from non-WEIRD countries and developing recommendations based on wider demographics [49]. We utilize this strategy through involving participants from non-WEIRD and WEIRD countries, investigating their
perceptions of agents, and asking them how they envision their “ideal conversational agents”. Furthermore, we develop design recommendations to align agents with various audiences’ desires.

Another strategy to reduce the gap between non-WEIRD- and WEIRD-centric technology is to empower those from non-WEIRD countries to develop their own technology. There are a number of tools that help enable nearly anyone to develop technology, many of which utilize visual or block-based coding [22]. These tools have largely been born out of the constructionist movement in education, which encourages the use of low-floor, high-ceiling programming tools to empower a wide variety of people to learn to program, including other underrepresented groups in the technology sector, like children [43]. Scratch, for instance, allows children to program their own web-based animations using block-based coding [44]. Other low-floor platforms enable users to develop conversational agents, including the Flow Editor and Alexa Blueprints [1, 24]. The MIT App Inventor platform allows users to develop fully-fledged apps, which can be deployed to mobile devices’ app stores [61], as well as conversational agents, which can be deployed to Amazon Alexa devices, through “ConvoBlocks” [34, 52, 53, 57].

In addition to developing design recommendations in this paper, we also utilize educational interventions to further democratize conversational agent technology. These interventions empower students—children and parents from WEIRD and non-WEIRD countries—to develop their own agents, while empowering us to understand how such participants envision and design agents. We adopted ConvoBlocks for agent development in our study, as it is open-source and allows nearly anyone to create deployable agents [52, 53]. Through constructionist workshops with this tool, we inform participants about how agents work and technology’s societal impact, while investigating how their partner models and trust change. The results informed the development of agent design recommendations.

1.1 Research Questions

Through engaging children and parents from non-WEIRD and WEIRD countries in conversational agent and societal impact curriculum, including agent-development, learning and design sessions, we aimed to answer the following research questions:

**RQ1:** How do people of various backgrounds (WEIRD and non-WEIRD, as well as different generations) perceive Alexa with respect to partner models [15] and trust before, during and after conversational agent development and societal impact activities?

**RQ2:** How do people of various backgrounds envision the future of conversational agents?

The results from these questions led to design recommendations for conversational agents based on user’s perceptions, and recommendations to align conversational agent development with users’ ideal future conversational agents. We also developed recommendations for agent pedagogy, which we present in our other paper [57].

2 BACKGROUND AND RELATED WORK

2.1 Trust of Conversational Agents

Because conversation is one of the most intuitive, primary methods humans use to communicate with each other, conversational interfaces are uniquely positioned to inspire relational interactions with technology [40, 47]. For instance, an agent recently won a Peabody Award for engaging in “emotional interactions, empathy, and connection” [11]. Furthermore, researchers have found correlations between human-agent relationship development and increased trust of agents [47]. Considering how trust is a key factor in misinformation spread [46, 62], we decided to specifically investigate people’s trust of agent’s correctness in this study. We also chose to emphasize children’s trust in this study,
as the risks associated with misinformation spread could be particularly acute with children (since they do not have the same critical analysis skills as adults).

Other studies have investigated people’s trust of conversational agents’ correctness. One example includes a study in which clinicians decide whether or not to utilize agents’ advice on diagnoses [18]; another includes a study in which customers decide whether or not to follow agents’ recommendations [32]. Nonetheless, few studies have investigated children’s or those from non-WEIRD countries’ trust of agents [17]. Even fewer have investigated how this trust may change through educational interventions. One example includes a study in which children engage in social robot curriculum, including modules on conversational AI, computer vision and societal impact, among others. If participants engaged in the societal impact module, their trust of the robot generally decreased [14]. Another example includes a study with ConvoBlocks in which students engaged in curriculum entirely focused on conversational agents, including their societal impact. In this study, researchers did not find any significant differences in trust through the curriculum. They did, however, observe concerning correlations between children’s perceived friendliness and trust of agents [56]. In both of these studies, however, the researchers only investigated general trust.

Many researchers have developed methods to investigate specific aspects of trust, such that developers can better assess which aspects of their technology affect such trust [9]. In our study, we adopt McKnight and Chervany’s widely-used model, which has four main components: (1) competence, (2) benevolence, (3) integrity and (4) predictability [33]. In our study, we found participants most often referred to predictability when discussing trust. Based on this and other results, we develop design recommendations, including the suggestion to develop agents to portray accurate levels of predictability to foster appropriate levels of trust.

2.2 Other Perceptions of Conversational Agents

People’s partner models, or mental models of their conversational partners, can significantly affect how they interact with agents. For instance, researchers have found that people make different language choices depending on their initial expectations of partner models [12, 15]. Partner models can be described in terms of three main dimensions: (1) competence and dependability, (2) human-likeness, and (3) cognitive flexibility [12, 15]. Designing agents that produce partner models that align with the capabilities of the agent (e.g., producing a partner model of perceived limited flexibility, if the agent is truly limited in flexibility), could help minimize user frustrations and ease conversation [15]. However, a deep understanding of conversational agent users’ partner models—and especially children’s partner models—is not reflected in the literature [15, 17].

Certain studies have investigated children’s general perceptions of conversational agents. For instance, one study found that the majority of 5-6 year old children considered agents to be friendly, alive, trustworthy, safe, funny, and intelligent [31]. Another study investigated 3-10 year old children’s perceptions, and found that children had different perceptions of agents’ intelligence depending on the modality of interaction with conversational agents. Others found students perceived agents to be more intelligent and felt closer to them after learning to program them [56]. None of these studies specifically investigated children’s partner models of agents.

2.3 Design Guidelines

With the recent major advances in conversational agent and voice-first systems, researchers have described a need for guidelines similar to the ones we have for graphical user interfaces (e.g., [41, 42]), but for voice user interfaces, such that designers can create effective, usable conversational agents. In the past few years, many different researchers have developed such agent-specific guidelines [37, 38]. For instance, researchers recently developed voice user interface
heuristics specifically for graphical user interface-trained designers, such that they can easily shift to design with the new interaction mode [38]. Other researchers developed voice user interface guidelines through an analysis of the Star Trek agent, “Computer” [3]. Others utilized interviews and relevant theories, like self-determination theory or values-based design, to develop conversational agent design guidelines [10, 63].

As a basis for our design guideline development, we adopt Murad et al. [36]’s guidelines, which address the criteria in the appendix [54]. Among the criteria, the guidelines had to be general enough to apply to agents with many different purposes, had to have specific evaluative usability recommendations, and had to have been used by many others in the literature [55]. Murad et al.’s guidelines are based on prominent usability guidelines for graphical user interfaces that developers have relied on for many years [41, 42, 48]. Murad et al. adapted the graphical guidelines to be agent-specific, and added two other agent-specific guidelines that address the unique challenges of transparency and speech-context of conversational agent design. They authors also mention how the human-computer interaction community needs to continue to explore agent design to refine the usability of such systems.

To continue this exploration, we investigate perspectives on the future of technology from underrepresented groups, including those of children or of those from non-WEIRD countries. We also investigate such groups’ perceptions of agents and develop guidelines to address these perceptions. Through this research, we aim to increase the diversity of perspectives in conversational agent design. We also encourage other researchers continue in this vein, as the world is full of many perspectives not yet observed.

3 LIMITATIONS AND FUTURE WORK

In this paper, we focus on voice-based agents due to humans’ long history of voice-based interactions and how this mode of interaction may cause agents to seem especially personified (and likely especially trustworthy [47, 56]). Nonetheless, future research may investigate people’s perceptions of text-based agents, as they are also common and have great potential for societal impact. Since we specifically used the voice-based agent of Amazon Alexa (as this is the only current type of agent the ConvBlocks platform supports [34]), its default persona could have biased people’s perceptions of agents. Future research could investigate how developing agents with different voices and on different platforms affects perceptions.

Another limitation includes how—although the number of WEIRD and non-WEIRD participants were balanced overall—the majority (57%) of participants from WEIRD countries were from the US, and the majority (87%) of participants from non-WEIRD countries were from Indonesia. Also, there were more children (59%) overall than parents. Future research could include more participants and further balance them across groups to verify results.

One other limitation includes the context of the study. Since the participants engaged in the workshops over Zoom, other factors in their environment could have affected the results. Another limitation includes how we leave the definition of “accurate” partner models and “appropriate” levels of trust to future research, and only investigate how participants’ perceptions of these change in our study.

Despite these limitations, this paper successfully presents five conversational agent design recommendations based on an initial investigation into conversational agent perspectives from non-WEIRD and WEIRD participants. The authors encourage future research to continue to build on such contributions to develop even more robust diverse perspectives.
4 PARTICIPANTS

Study participants came from various backgrounds (non-WEIRD and WEIRD countries), various generations (children and parents), and various prior experiences (e.g., programming, AI and conversational agent experience). Interest forms for the study were sent to educational email lists worldwide, attracting 99 pairs of interested children and parents. Forty-nine participants completed research consent forms, and completed at least 1 of the 3 surveys that were given before (46), during (40), and after (35) the study. According to the pre-survey, there were 27 children (age average=13.96, SD=1.829), 19 parents (age average=46.35, sd=11.07), 23 WEIRD (age average=26.45, SD=19.24), and 23 non-WEIRD (age average=25.48, SD=15.18) participants. There were four WEIRD countries (the U.S, Singapore, Canada and New Zealand) represented, with the majority of the WEIRD participants from the U.S. There were four non-WEIRD countries (Indonesia, Iran, Japan, and India) represented, with the majority of the non-WEIRD participants from Indonesia.

Other statistics include how twenty participants identified as female, 25 identified as male, and 1 identified as non-binary. Fourteen participants had no prior programming experience, 6 only had visual (or blocks-based) programming experience, and 26 had text-based programming experience. Thirty-eight participants reported typically using conversational agents in their first language; 8 reported typically using them in another language.

5 PROCEDURE

5.1 The ConvoBlocks Platform

ConvoBlocks is an open-source, block-based programming platform within the App Inventor environment, which allows nearly anyone to program conversational agents [34, 52, 61]. To do so, students first define their agent’s invocation name (e.g., “My Carbon Footprint Agent”), intents (e.g., groups of phrases like, “Calculate my carbon footprint”, “What’s my carbon footprint?”, etc.) and entities (e.g., information units like number of miles driven, kilowatts of energy used, etc.) the agent should be able to recognize. Through the process of agent development, students learn conversational agent terminology and concepts, which are described in-detail in the appendix [54]. Next, students define how the agent responds to the defined intents (e.g., “You have a carbon footprint of 11 tonnes/year”). They can do so using the web pages shown in Figure 2. After this, students can test their agent on ConvoBlocks, or deploy their agents to any Alexa-enabled devices, like the Alexa mobile app or an Echo Spot [53].

Fig. 2. Two web pages from ConvoBlocks [34], allowing users to define invocation names, intents and entities, and then program agents’ responses to intents.
5.2 Workshops

The curriculum consisted of two days of 3.5-hour Zoom classes about conversational agents. We repeated the workshops twice, once with participants from WEIRD countries and once with participants from non-WEIRD countries. For each repetition, the first day of the curriculum focused on programming agents and the second day focused on societal impact activities. On the first day, we provided a pre-survey to all participants with questions about their demographics, trust of conversational agents, and self-identification as programmers. Afterwards, instructors led participants through two conversational agent development tutorials, as described in the thesis, [55]. We gave participants written versions of these two tutorials, plus a third, such that they could complete them at their own pace, and attempt the third if they finished early. These tutorials focused on teaching conversational agent concepts described in the appendix [54] and another paper [57] (which focuses on the conversational agent pedagogy of this study). The tutorials taught participants to create agents that responded to questions about carbon footprints, as shown in Figure 3.

![Figure 3](image)

Fig. 3. An example conversation with the agent developed in the workshop tutorials.

In between the two tutorials, the group completed an ideation design session where they responded to prompts about what their “ideal” agent would look like, sound like, do, and say (among other prompts) using a virtual whiteboard with sections for each major subset (i.e., children from non-WEIRD countries, parents from non-WEIRD countries, children from WEIRD countries, etc.). A portion of the section for children from non-WEIRD countries is shown in Figure 1. After the tutorials and design session, we provided participants with a “mid-survey”. This survey contained the same questions as the first survey, along with further questions asking if students’ opinions had changed. The second day included presentations and group discussions about societal impact of technology. Participants gathered in groups of 2-4 children with their parents for the discussions. The presentations encouraged participants to think about the impact (positive and negative) of technology; the discussions explored how technology could help address world problems, like sustainability, with an emphasis on conversational agents as part of the solution. In the final activity, small groups of participants presented on their proposed solutions to the entire group. They had the opportunity to
design conversational agents, which they could demonstrate in their presentations. We provided the post-survey to participants, which asked similar questions to the first two surveys, after the presentations.

5.3 Data analysis

The pre-, mid- and post-surveys included Likert scale and short answer questions. For example, we asked students to respond to the prompt “Conversational agents (e.g., Siri, Alexa, Google Home) say things that are...” using a 5-point scale from “Always Right” to “Always Wrong”. We also asked students to write sentence responses to questions like, “Please explain why you think conversational agents say things that are right/wrong”. To analyze the Likert scale data, we used Mann-Whitney U tests, Wilcoxon signed-rank tests, and independent and paired t-tests, depending on the sample and distribution of the data. We identify statistical significance in Figures using star symbols (i.e., “*” for $p \leq .05$, “**” for $p \leq .01$ and “***” for $p \leq .001$).

To analyze the responses to the short-answer questions as well as the prompts during the design session, we used a coding reliability approach to thematic analysis [7]. Three researchers tagged each section of the data and reconvened to agree on common sets of themes, including guidelines and definitions for each theme. The theme definitions are shown in the appendix [54]. The researchers completed three rounds of coding such that the Krippendorff’s Alpha between all researchers was $\alpha \geq .800$ [2]. We aggregated the tagged data by union between researchers, and organized them with respect to the categories, WEIRD, non-WEIRD, child and parent.

6 RESULTS

This section describes the results most relevant to agent design recommendations. We describe other results (e.g., most relevant to pedagogy recommendations) in [55, 57].

6.1 All Participants

6.1.1 Partner Model.

Sixty-two percent of overall participants indicated they felt their partner models changed through the programming activity in their long-answer responses, as shown in Table 1. Although there were few significant differences in the overall Likert scale data, one includes how after the workshops participants thought of agents as more of friends than co-workers (pre/post: $\bar{x}=3.58, 3.24$; $t(32)=2.15$; $p=.039$).

| Subset          | Changed | Did not change | Ambiguous |
|-----------------|---------|----------------|-----------|
| Overall participants | 62%     | 35%            | 3%        |
| Non-WEIRD       | 76%     | 18%            | 6%        |
| WEIRD           | 50%     | 50%            | 0%        |
| Children        | 67%     | 33%            | 0%        |
| Parents         | 54%     | 38%            | 8%        |

6.1.2 Trust.

Participants overall (and all major subsets too, including those from WEIRD and non-WEIRD countries, children, and
Table 2. Percent of long-answer responses indicating participants generally trusted, distrusted, or were unsure about whether they trusted conversational agents on the pre-survey.

| Subset        | Trustworthy | Not trustworthy | Unsure |
|---------------|-------------|----------------|-------|
| Overall       | 19%         | 72%            | 9%    |
| Non-WEIRD     | 15%         | 70%            | 15%   |
| WEIRD         | 22%         | 74%            | 4%    |
| Children      | 20%         | 72%            | 8%    |
| Parents       | 17%         | 72%            | 11%   |

parents) prior to, during and after the workshops, generally trusted Google, Alexa and newspapers significantly more than both parents and friends to report correct information. Figure 4 shows this trend. In other words, people tended to trust technology more than people, and their parents more than friends for correct information.

In a long-answer question, we asked participants why they trusted or distrusted agents to provide correct information, to which most provided answers for why they distrusted agents. Table 2 shows the results from before the programming activity, and Table 3 shows the results from after the programming activity. After the programming activity, the percent of answers indicating distrust decreased slightly for all subsets.

We found participants’ reasoning for trust leaned towards the aspect of competence for nearly all subsets on both the pre- (Table 4) and mid-survey (Table 5). We found no responses indicating participants considered the benevolence aspect of trust with respect to conversational agents.

Figure 5 compares themes in participants’ reasoning for their opinions on trust before and after the programming activity. Both before and after the activity, participants most often mentioned the source of the agent’s information,
Table 3. Percent of long-answer responses indicating participants’ trust of conversational agents on the mid-survey.

| Subset            | Trustworthy | Not trustworthy | Unsure |
|-------------------|-------------|-----------------|-------|
| Overall participants | 22%         | 64%             | 14%   |
| Non-WEIRD         | 13%         | 67%             | 20%   |
| WEIRD             | 29%         | 62%             | 10%   |
| Children          | 27%         | 59%             | 14%   |
| Parents           | 14%         | 71%             | 14%   |

Table 4. Percent of long-answer responses indicating different aspects of McKnight and Chervany’s trust model when participants discussed their opinions on trust of conversational agents on the pre-survey.

| Subset | Competence | Integrity | Predictability | Benevolence |
|--------|------------|-----------|----------------|-------------|
| Overall | 39%        | 25%       | 36%            | 0%          |
| Non-WEIRD   | 36%        | 29%       | 36%            | 0%          |
| WEIRD      | 41%        | 23%       | 36%            | 0%          |
| Child      | 34%        | 30%       | 36%            | 0%          |
| Parent     | 48%        | 17%       | 35%            | 0%          |

Table 5. Percent of long-answer responses indicating different aspects of McKnight and Chervany’s trust model when participants discussed their opinions on trust of conversational agents on the mid-survey.

| Subset | Competence | Integrity | Predictability | Benevolence |
|--------|------------|-----------|----------------|-------------|
| Overall | 43%        | 23%       | 34%            | 0%          |
| Non-WEIRD   | 37%        | 22%       | 41%            | 0%          |
| WEIRD      | 47%        | 24%       | 29%            | 0%          |
| Child      | 37%        | 26%       | 37%            | 0%          |
| Parent     | 52%        | 17%       | 30%            | 0%          |

including human data, the internet and other sources, as reasons for their opinions on trust. (See the appendix [54] for descriptions of the themes.)

![Overall Reasoning for Trust Levels](image)

Fig. 5. Overall participants’ responses to the question asking about their reasoning for their opinions on trust of agents in terms of counted tag frequency. (See the appendix [54] for descriptions.)
6.1.3 Ideal Agents.
Participants described their ideal conversational agents with more task-oriented (75%) than non-task oriented (or socially-oriented; 25%) language, as shown in Figure 6. (See the appendix [54] for example task and non-task oriented phrases.) All subsets of participants analysed (e.g., children, parents, those from WEIRD countries, etc.) also showed this task-orientation, albeit with different proportions. Participants also used slightly more human-like (55%) than artificial (45%) descriptions (as shown in 7) for their agents. However, this was not the case for all subsets of participants, including how children from non-WEIRD countries described their ideal agents with more artificial descriptions (59%). (See the appendix [54] for example human-like and artificial-related phrases.)

![Preferences for Task-Orientation of Agents](image)

Fig. 6. The number of phrases indicating a preference for either task-oriented or non-task oriented (i.e., socially-oriented) agents normalized and grouped by various subsets of the participants. Reproduced from the thesis, [55].

The frequency of themes describing ideal agents from most to least frequent are shown in Figure 8. The descriptions of these themes can be found in the appendix [54].

6.2 Participants from WEIRD vs. Non-WEIRD Countries

6.2.1 Partner Model.
After the programming activity, there was a significant difference between how participants from WEIRD and non-WEIRD countries felt about Alexa’s competence ($\bar{x}=3.00,2.11; U(38)=106.5; p=.004$), as shown in Figure 9. Participants from WEIRD countries thought Alexa was less competent after the programming activity than before (pre/mid: $\bar{x}=2.43,2.95; t(20)=-2.33; p=.030$). Children from non-WEIRD countries thought Alexa was more competent after the workshops than before (pre/post: $\bar{x}=3.20,2.30; W(9)=0; p=.0067$).

After the programming activity, participants from WEIRD countries thought Alexa was more of a peer than an authority figure than what participants from non-WEIRD thought ($\bar{x}=3.64,3.00; U(38)=134; p=.036$), as shown in Figure 10. Those from WEIRD countries thought Alexa was less dependable than what those from non-WEIRD countries thought ($\bar{x}=3.23,3.78; U(38)=122.5; p=.014$), as shown in Figure 11. This may have been due to how participants from non-WEIRD countries’ perspectives shifted through the programming activity, as they thought Alexa seemed more
dependable afterwards than before (pre/mid: $\bar{x}=3.39,3.78; W(17)=0; p=.020$). Participants from non-WEIRD countries also thought Alexa seemed more flexible afterwards than before (pre/mid: $\bar{x}=2.89,3.33; W(17)=4.5; p=.021$).

6.2.2 Trust.

Participants from WEIRD countries trusted Alexa (pre/post: $\bar{x}=4.00,3.75; W(15)=0; p=.046$) less to give correct information after the workshops than before. There was no significant difference between pre- and post-survey trust results for those from non-WEIRD countries.

6.2.3 Ideal Agents.

As shown in Figure 6, when commenting on ideal conversational agents, participants from WEIRD countries had relatively more task-orientation (80%) than participants from non-WEIRD countries (70%). Participants from non-WEIRD countries commented relatively more on how conversational agents should be artificial (55%) than those from WEIRD countries (27%) (as shown in Figure 7).

As shown in Figure 8, people from WEIRD countries tended to focus more on basic (already common conversational agent) features and pop-culture or familiar features than those from non-WEIRD countries, whereas people from non-WEIRD countries tended to focus more on proactivity, cultural intelligence and addressing concerns than those from WEIRD countries.

6.3 Parents vs. Children

6.3.1 Partner Model.

Before ($\bar{x}=2.74,2.11; U(44)=167; p=.018$) and after ($\bar{x}=2.79,2.13; U(38)=112; p=.0093$) the programming activity, children thought Alexa was more human-like than parents did. They also thought Alexa was warmer than their parents did before ($\bar{x}=2.70,3.37; U(44)=170.5; p=.021$), during ($\bar{x}=2.96,3.56; U(38)=129.5; p=.034$) and after ($\bar{x}=2.62,3.50; U(33)=81.5; p=.011$) the workshops. After the programming activity, they thought Alexa was more dependable than their parents did ($\bar{x}=3.82,3.14; U(16)=21; p=.039$).
Fig. 8. Bar charts showing the relative frequency of phrases tagged with particular themes for overall participants (top), parents vs. children (middle), and WEIRD vs. non-WEIRD (bottom). Reproduced from the thesis, [55].

Fig. 9. Participants from non-WEIRD and WEIRD countries’ responses when asked to rate Alexa’s competence on a 5-point Likert scale question given after the programming activity. Reproduced from the thesis, [55].
6.3.2 Trust.
As shown in Figure 12, after the programming activity, children trusted Alexa to be more correct than parents did ($\bar{x}=4.04, 3.63; U(38)=127.5; p=.023$). Children also trusted agents to report correct information more after the societal impact activity than before (mid/post: $\bar{x}=2.60, 2.35; t(19)=2.52; p=.021$).

6.3.3 Ideal Agents.
As shown in Figure 6, when commenting on their ideal conversational agents, parents had relatively more task-orientation (82%) than children (71%). Children commented relatively more on how conversational agents should be artificial (52%) than parents did (30%) (as shown in Figure 7). As shown in Figure 8, parents tended to focus more on
personalized features and pop-culture or familiar features than children, whereas children tended to focus more on fun features, approachable/friendly features, and addressing concerns than parents.

6.4 Participants of Different Gender Identities

6.4.1 Partner Model.
Male participants’ opinion of Alexa’s interactivity (pre/post: $\bar{x}=2.84,3.42; W(18)=6; p=.024$) and companionship (pre/post: $\bar{x}=3.74,3.26; W(18)=8; p=.039$) changed through the workshops, as they felt Alexa was less interactive and more like a friend after the workshops than they did before. There were no significant differences in female participants’ opinions overall in terms of the partner model through the workshops.

6.4.2 Trust.
For those from WEIRD countries, prior to the workshops, female participants thought Alexa reported less correct information than male participants did ($\bar{x}=3.69,4.18; U(22)=42.5; p=.031$) and thought their friends reported more correct information than male participants did ($\bar{x}=3.31,2.91; U(22)=47; p=.041$). These differences were not found for males vs. females from non-WEIRD countries. Figure 13 shows the distribution of the results in terms of trust of Alexa’s correctness.

![WEIRD Participants’ Trust of Alexa’s Correctness](image)

Fig. 13. Female and male participants from WEIRD countries’ responses when asked to rate their trust of Alexa’s correctness on a 5-point Likert scale question given before the programming activity. Reproduced from the thesis, [55].

6.5 Participants with Different Levels of Prior Programming Experience

6.5.1 Partner Model.
Compared to participants who had no programming experience, those who had text-based programming experience thought Alexa was less competent ($\bar{x}=2.73,2.07; W(16)=0; p=.038$) before the workshop.

6.6 Participants with Different Experiences with Conversational Agents

6.6.1 Partner Model.
At all times throughout the workshop, participants who used conversational agents in their first language thought Alexa was more human-like than those who used them in another language. Before the workshop activities, they also thought Alexa was more correct than those who used it in another language ($\bar{x}=4.03,3.00; U(44)=52; p=5.50 \times 10^{-4}$).
7 DISCUSSION

7.1 RQ1: Partner Models and Trust

This section develops design recommendations for conversational agents (labelled with “P”) based on participants’ perceptions of partner models and trust of agents.

7.1.1 P1: Inform users about trustworthiness.

In general, participants trusted Alexa more than their parents or friends to give correct information. This may indicate an over-trust of Alexa, depending on the actual correctness of the device (although we leave this as a question for future research). Since different agents show varying levels of correctness [30], different agents should be trusted differently. To foster levels of trust matching agents’ actual trustworthiness, developers can design agents to indicate their accuracy. For instance, agents may explicitly mention that their information source might be incorrect, or use diction to indicate uncertainty, like “might”, “perhaps” or “probably”. One real-world example includes how the Google Assistant often reports the source of its information, or how Alexa often explicitly states, “Sorry, I don’t know that”. Future areas of research include determining how such phrases and diction affect users’ partner models and feelings of trust towards agents, and quantifying appropriate levels of trust for various agents.

In order to better inform users and allow them to develop more accurate understandings of agents, we propose the design recommendation, P1: Inform users about trustworthiness, as shown in Table 6. This may include explicit or implicit phrases and diction to indicate how confident the agent is about the accuracy or helpfulness of its responses. It also may focus on the aspects of agents’ competence, then predictability and then integrity, as participants referenced these aspects most often when describing their trust of agents. More specifically, in their descriptions, participants often referenced the programming of the agent; whether or not agents understood phrases correctly; and where agents obtained their information. Thus, agent designers may want to be transparent about these areas in their designs.

Table 6. Conversational agent design recommendations with respect to different subsets of participants’ partner models and trust of agents.

| Perceptions-based design recommendation | Opportunities in practice |
|----------------------------------------|---------------------------|

16
**P1: Inform users about trustworthiness**

- Develop conversational agents to include explicit or implicit diction to indicate how confident the agent is about the accuracy of its responses (e.g., “probably”, “unsure”, “likely not”, etc.)
- Develop conversational agents with the ability to explain themselves and provide transparency in terms of their abilities to provide trustworthy information [56]
- Focus on the competence, then predictability and then integrity aspects of trust when developing agents; for instance, by informing end users how the agent obtains its information, what it does with information given to it and how it is programmed
- Since personification may increase users’ trust of agents, align the amount of personification of agents with the actual trustworthiness of the conversational agent [56] (see also, P2)
- For audiences that may particularly find conversational agents more trustworthy through increased interaction (e.g., children) ensure conversational agent development emphasizes trust indicators throughout interactions
- For audiences that may have more initial distrust of agents (e.g., females from WEIRD countries), developers may want to use techniques to encourage trust, such as developing trustworthy personas (see P2)

**P2: Design conversational agent personas to foster appropriate partner models**

- Design conversational agent personas to foster appropriate relationship building (e.g., whether that means shifting perceptions from co-worker to friend or vice-versa) and therefore trust, as described in [47] (see also, P1)
- Develop diverse, multilingual conversational agent personas that many people can relate to and understand (see Section 7.1.2 about similarity and familiarity of conversational agents)
- For an audience from non-WEIRD countries, developers may want to focus on creating competent personas
- For an adult audience, developers may want to focus on creating more human-like and warm personas
- For an audience with lack of trust in agents’ correctness, developers may want to ensure their conversational agents use the audience’s first language

There are similarities between P1 and two of Murad et al.’s design recommendations, G1 (Visibility/Feedback of System Status) and A1 (Ensure Transparency/Privacy). (See the appendix [54] for a list of all recommendations). All three recommendations focus on some aspect of informing the user about the agent; however, the particular information each recommendation emphasizes is unique. G1 focuses on informing users about how agents work such that they can effectively complete their goals, A1 focuses on informing users about data collection and their privacy when using agents, and P1 focuses on informing users about agents’ accuracy and trustworthiness.

P1 is also supported by the results from Van Brummelen et al.’s study with children and ConvoBlocks. The authors found correlations between conversational agents’ friendliness and children’s perceived trustworthiness, and described the importance of ensuring children’s feelings of trust are proportionate to the actual trustworthiness of the device.
They also described how developers can create agents with the ability to explain their limitations, which can foster users’ trust to appropriate levels [56].

This trust- and relationship-building effect is not only limited to children working with ConvoBlocks. For example, for all major subsets, we found there were fewer responses mentioning distrust of agents after the programming activity (see Table 2 and 3). We also found overall participants’ (including parents’) feelings shifted towards Alexa being more of a friend than a co-worker through the workshops. Another study linked relationship-building with agents to increased trust [47]. This presents a design challenge, as increased trust has been linked to increased misinformation spread [46, 62]. Furthermore, another study found adults described trustworthiness as a key component in conversation, and emphasized how agents should indicate their trustworthiness [10].

It is also important to consider various audiences and their experiences over time when designing for trust, since different aspects of the workshops affected different subsets of the participants’ trust differently. For instance, participants from WEIRD countries’ feelings of trust towards Alexa decreased through the workshops, but there were no significant differences for participants from non-WEIRD countries in this way. Nonetheless, for the subset of children from non-WEIRD countries, there was a significant difference. Children thought agents were more trustworthy after the programming activity, perhaps because of increased relationship (as described in [47]) or the Dunning-Kruger effect (as described in [56]). In general, children trusted Alexa to be more correct than parents did after the programming activity. Children’s perceptions of correctness increased after the societal impact activity. There were also differences in trust for different genders. For instance, female participants from WEIRD countries tended to distrust agents more, and trust their friends more than male participants from WEIRD countries did. Table 6 lists further design implications of these results.

7.1.2 P2: Design conversational agent personas to foster appropriate partner models.

At various times in the workshops, there were significant differences between subsets of participants’ partner models. For instance, after the programming activity, participants from WEIRD countries felt Alexa was less competent, less dependable and more of a peer than an authority figure than those from non-WEIRD countries did. Children felt Alexa was warmer, more human-like, and more dependable than parents did at various times during the workshops. Those with text-based programming experience thought Alexa was less competent than those who had no programming experience prior to the workshops.

Furthermore, subsets of the participants’ perceptions of Alexa changed differently through the activities. For instance, before and after the programming activity, children thought Alexa was more human-like than parents did; however, after the societal impact activity, there was no significant difference. After the programming activity, children thought Alexa was more dependable than their parents did, but not before. Male participants’ opinion of Alexa’s interactivity and companionship changed through the workshops; however, there were no significant differences in female participants’ opinions overall in terms of the partner model through the workshops.

Thus, different activities may change certain people’s perceptions of conversational agent partner models, but not others; and certain people may have particular initial perceptions that others do not share. This could have implications for how people of different backgrounds and ages interact with Alexa. According to Hinds et al., team members or partners are chosen according to their reputation for being competent, strength of prior relationship, and perceived similarity [21]. In this context, users may interact with agents as conversational partners more often if the agents are perceived as competent, friends, and similar to users. Other studies have emphasized the importance of providing agents with distinct personalities [26, 59], which could be modeled by partner models [15]. Others found users prefer agents
similar to themselves [39, 50]. This leads to another recommendation, **P2: Design conversational agent personas to foster appropriate partner models**, as described in Table 6.

This recommendation could vary greatly depending on the intended agent audience. For instance, since participants from WEIRD countries felt Alexa was less competent and dependable than those from non-WEIRD countries, perhaps agent designs for audiences in non-WEIRD countries should focus on increasing this perception of competence. Since those who used agents in their first language thought Alexa was more human-like before, during and after the workshops than those who used them in another language, and Hinds et al. note that partners are often chosen based on similarity or familiarity, perhaps agent designers should focus on developing multilingual and diverse agent personas depending on their intended audience [21]. Table 6 includes design implications, such as these, with respect to different audiences.

### 7.2 RQ2: Envisioning Future Conversational Agents

This section discusses how people of various backgrounds envision their ideal future conversational agents, and develops recommendations for aligning designs with these ideal future agents (labeled with “F”).

#### 7.2.1 F1: Consider the degree of task-orientation of agents appropriate to end-users’ cultural perspectives.

All major subsets of participants (i.e., children, parents, participants from non-WEIRD countries and participants from WEIRD countries) described their ideal agents with more task orientation than social orientation. Participants’ perspectives may have been influenced by how current agents tend to be task-oriented, rather than truly conversational or social [10]. That said, participants also included social (non-task) oriented agent attributes in their responses, like having agents ask about how users feel. There were also different ratios of task vs. social orientations depending on the subset. For instance, children’s responses indicated a desired social orientation of 25%, whereas parents’ responses indicated 18%. Participants from non-WEIRD countries’ responses indicated a desired social orientation of 30% whereas participants from WEIRD countries’ responses indicated 20%. Thus, we suggest developing agents with varying amounts of task-orientation with respect to the intended users, **F1: Consider the degree of task-orientation of agents appropriate to end-users’ cultural perspectives**, as explained in Table 7.

#### 7.2.2 F2: Consider the degree of agents’ human-likeness with respect to end-users’ cultural perspectives.

Participants also commented on agents’ human-likeness and artificiality, with statements such as, “[My ideal agent would be] like a robot, but not human like otherwise it would be a bit creepy”, as shown in the appendix [54]. Overall, they described their ideal conversational agents with a slight preference for human-like agents over artificial ones. This varied greatly between the major subsets of participants, as shown in Figure 7. For instance, children described a slight preference for artificial agents (52%), whereas parents described a large preference for human-like agents (70%). Similarly, those from non-WEIRD countries described a slight preference for artificial agents (55%), whereas those from WEIRD countries described a large preference for human-like agents (73%). Thus, we propose the design recommendation, **F2: Consider the degree of agents’ human-likeness with respect to end-users’ cultural perspectives**, as shown in Table 7. This aligns with human-robot interaction research around the “uncanny valley”, which describes how—for humans to have affinity for a robot—it is undesirable for it to be completely human-like or completely artificial [35].

#### 7.2.3 F3: Focus development on useful, common features; user-orientation; enjoyable interactions; and emotional intelligence with respect to end-users’ cultural perspectives.

We identified nine themes (other than task-orientation and human-likeness) participants desired in their agents. Three of the themes indicate participants want future conversational agents to be user-oriented (Convenient, Personalized,
and Proactive); three indicate a desire for enjoyable interactions (Approachable/friendly, Familiar or pop-culture related, and Fun); and two indicate a desire for emotional intelligence (Addresses concerns and Culturally intelligent). The final theme, Basic features, indicates participants want future agents to include the typical features current agents have, like the ability to play music or get the weather.

In terms of relative importance to participants overall, they mentioned Basic features most often, then user-oriented themes, then enjoyable interactions, and finally emotional intelligence. In terms of theme rankings for different subsets, both parents and those from WEIRD countries emphasized pop-culture references while de-emphasizing addressing concerns. From the relative differences shown in Figure 8, those from non-WEIRD countries emphasized cultural intelligence, parents emphasized personalization, and children emphasized friendliness and approachability. Thus, we propose the design recommendation, **F3: Focus development on useful, common features; user-orientation; enjoyable interactions; and emotional intelligence with respect to end-users’ cultural perspectives**, as described in Table 7. This is also supported by the results from Van Brummelen et al.’s study with ConvoBlocks, in which students reported most often using basic, useful features when interacting with Alexa [56].

Table 7. Design recommendations for future conversational agents with respect to how different subsets of participants described their ideal conversational agents.

| Design recommendation for the future of conversational agents | Subset-specific information                                                                 |
|---------------------------------------------------------------|---------------------------------------------------------------------------------------------|
| **F1:** Consider the degree of task-orientation of agents appropriate to end-users’ cultural perspectives | **Children:** Relatively more non-task orientation (29%) than overall participants  
**Parents:** Relatively less non-task orientation (18%) than overall participants  
**Non-WEIRD:** Relatively more non-task orientation (30%) than overall participants  
**WEIRD:** Relatively less non-task orientation (20%) than overall participants |
| **F2:** Consider the degree of agents’ human-likeness with respect to end-users’ cultural perspectives | **Children:** Larger preference for artificial (52%) than overall participants  
**Parents:** Larger preference for human-likeness (70%) than overall participants  
**Non-WEIRD:** Larger preference for artificial (55%) than overall participants  
**WEIRD:** Larger preference for human-likeness (73%) than overall participants |
F3: Focus development on useful, common features; user-orientation; enjoyable interactions; and emotional intelligence with respect to end-users’ cultural perspectives

| Children: More emphasis on the Addresses concerns and Approachable/friendly themes; and less emphasis on the Familiar or pop-culture and Personalized themes |
| Parents: More emphasis on the Familiar or pop-culture and Personalized themes; and less on the Addresses concerns and Approachable/friendly themes |
| Non-WEIRD: More emphasis on the Addresses concerns and Culturally intelligent themes; and less emphasis on the Familiar or pop-culture and Basic features themes |
| WEIRD: More emphasis on the Familiar or pop-culture and Basic features theme; and less on the Addresses concerns and Culturally intelligent themes |

8 CONCLUSIONS

This study investigated how people of various backgrounds (WEIRD and non-WEIRD, as well as different generations) perceive agents in terms of partner models and trust, and how they envision their ideal agents. The results led to two design recommendations based on users’ perceptions: P1: Inform users about trustworthiness and P2: Design conversational agent personas to foster appropriate partner models. It also led to three recommendations for how to align agent development with users’ ideal future agents: F1: Consider the degree of task-orientation of agents appropriate to end-users’ cultural perspectives, F2: Consider the degree of agents’ human-likeness with respect to end-users’ cultural perspectives and F3: Focus development on useful, common features; user-orientation; enjoyable interactions; and emotional intelligence with respect to end-users’ cultural perspectives. There are many opportunities to continue this research, as described in Section 3. We hope that through researchers’ continued work, and by developers’ utilization of our recommendations, we will increasingly design conversational agents “for all”.

ACKNOWLEDGMENTS

REFERENCES

[1] Amazon. 2021. Skill Blueprints. https://blueprints.amazon.com/. Accessed: 2021-09-14.
[2] Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34, 4 (12 2008), 555–596. https://doi.org/10.1162/coli.07-034-R2 arXiv:https://direct.mit.edu/coli/article-pdf/34/4/555/1808947/coli.07-034-r2.pdf
[3] Benett Axtell and Cosmin Munteanu. 2021. Tea, Earl Grey, Hot: Designing Speech Interactions from the Imagined Ideal of Star Trek. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 249, 14 pages.
[4] John Barryman. 2014. 18 Siri FAILs That Will Ruin Your Day. https://www.ranker.com/list/siri-fails-iphone-sucks/john-barryman. Accessed: 2021-09-01.
[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016), 4349–4357.
[6] Rishi Bommasani, Drew A Hudson, Ehsan Adel, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeanette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* 1 (2021), 214 pages.
[7] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic Analysis. In *Handbook of research methods in health social sciences*, Pranee Liamputtong (Ed.). Springer, Singapore, Singapore.
[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[9] Jin-Hee Cho, Kevin Chan, and Sibel Adali. 2015. A Survey on Trust Modeling. *ACM Comput. Surv.* 48, 2. Article 28 (oct 2015), 40 pages. https://doi.org/10.1145/2815595
[10] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garazulde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents.
Robert Holwerda and Felienne Hermans. 2018. A Usability Analysis of Blocks-based Programming Editors using Cognitive Dimensions. In

Ben Frederickson. 2018. Where Do The World’s Software Developers Live? https://www.benfrederickson.com/github-developer-locations/. Accessed:

Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers

Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, and Katharina Reinecke. 2021. How WEIRD is CHI?. In

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski,

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quaternity, Zion Mengesha, John R. Rickford, Dan Jurafsky, and

Hankyung Kim, Dong Yoon Koh, Gaeun Lee, Jung-Mi Park, and Youn-kyung Lim. 2019. Designing Personalities of Conversational Agents. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (Glasgow, Scotland Uk) (CHI EA ’19). Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312887

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quaternity, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117 arXiv:https://www.pnas.org/content/117/14/7684.full.pdf

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, and Emmanuel Dupoux. 2021. On Generative Spoken Language Modeling from Raw Audio. Transactions of the Association for Computational Linguistics 1, 1 (Feb, 2021), 20 pages. https://hal.inria.fr/hal-03329219

Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, and Katharina Reinecke. 2021. How WEIRD is CHI?. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. https://doi.org/10.1145/3411764.3445488

Gustavo López, Luis Quesada, and Luis A. Guerrero. 2018. Alexa vs. Siri vs. Cortana vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces. In Advances in Human Factors and Systems Interaction, Isabel L. Nunes (Ed.). Springer International Publishing, Cham, 241–250. https://doi.org/10.1007/978-3-319-78561-7_25

Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children’s Questions. In Proceedings of the 18th ACM International Conference on Interaction Design and Children (Boise, ID, USA) (IDC ’19). Association for Computing Machinery, New York, NY, USA, 301–313. https://doi.org/10.1145/3311927.3323150

Sin Man Lui and Wendy Hsu. 2010. Effects of Smiling and Gender on Trust Toward a Recommendation Agent. In 2010 International Conference on Cyberworlds. IEEE, Singapore, Singapore, 398–405. https://doi.org/10.1109/CW.2010.26
What Do WEIRD and Non-WEIRD Conversational Agent Users Want and Perceive? Towards Transparent, Trustworthy, Democratized Agents

D. Harrison McKnight and Norman L. Chernyav. 2001. What Trust Means in E-Commerce Customer Relationships: An Interdisciplinary Conceptual Typology. *International Journal of Electronic Commerce* 6, 2 (2001), 35–59. https://doi.org/10.1080/10864415.2001.11044235

MIT App Inventor. 2022. ConvolutionalBlocks. https://alexia.appinventor.mit.edu/. Accessed: 2022-09-14.

Masahiro Mori, Karl F. MacDorman, and Norri Kagcki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (2012), 98–100. https://doi.org/10.1109/MRA.2012.2192811

Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design Guidelines for Hands-Free Speech Interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) (MobileHCI ’18). Associations for Computing Machinery, New York, NY, USA, 269–276. https://doi.org/10.1145/3236112.3236149

Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2021. Finding a New Voice: Transitioning Designers from GUI to VUI Design. In *CHI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao [online], Spain) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 22, 12 pages. https://doi.org/10.1145/3469595.3469617

Clifford Nass and Kwan Min Lee. 2000. Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) (CHI ’00). Association for Computing Machinery, New York, NY, USA, 329–336. https://doi.org/10.1145/332040.332452

Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, Cambridge, MA.

Jakob Nielsen. 1994. Enhancing the Explanatory Power of Usability Heuristics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) (CHI ’94). Association for Computing Machinery, New York, NY, USA, 152–158. https://doi.org/10.1145/191666.191729

Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books, New York, NY.

Seymour A Papert. 2019. *Mindstorms: Children, computers, and powerful ideas*. Basic books, New York, NY.

Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, et al. 2009. Scratch: programming for all. Commun. ACM 52, 11 (2009), 60–67.

Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations. In *AICS, AICS, Ireland*, 104–115.

Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) (WebSci ’19). Association for Computing Machinery, New York, NY, USA, 265–274. https://doi.org/10.1145/3292522.3326012

William Seymour and Max Van Kleek. 2021. Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 371 (oct 2021), 16 pages. https://doi.org/10.1145/3479515

Ben. Shneiderman. 1998. *The design of everyday things: Revised and expanded edition*. Basic books, New York.

Seymour A Papert. 2020. *Mindstorms: Children, computers, and powerful ideas*. Basic books, New York, NY.

Mitchel Resnick, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay Silver, Brian Silverman, et al. 2009. Scratch: programming for all. Commun. ACM 52, 11 (2009), 60–67.

Elayne Ruane, Abeba Birhane, and Anthony Ventresque. 2019. Conversational AI: Social and Ethical Considerations. In *AICS, AICS, Ireland*, 104–115.

Christian Sturm, Alice Oh, Sebastain Linzen, Jose Abdelnour Noceara, Susan Dray, and Katharina Reinecke. 2015. How WEIRD is HCI? Extending HCI Principles to Other Countries and Cultures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (Seoul, Republic of Korea) (CHI EA ’15). Association for Computing Machinery, New York, NY, USA, 2425–2428. https://doi.org/10.1145/2702613.2702656

Adriana Tapus, Cristian Tăpău, and Maja J. Matarić. 2008. User–robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. *Intelligent Service Robotics* 1, 2 (01 Apr 2008), 169–183. https://doi.org/10.1007/s11370-008-0017-4

Reddit Users. 2015. Alexa misunderstandings... https://www.reddit.com/r/amazonecho/comments/41159u/alexa_misunderstandings/. Accessed: 2021-09-01.

Jessica Van Brummelen. 2019. Conversational Artificial Intelligence Intelligence Development Tools for K-12 Education. http://appinventor.mit.edu/papers/JessVBPublications/AAAI_Fall_Symposium_2019_SMConvAI_Final.pdf. In 2019 AAAI Fall Symposium. AAAI, Washington, DC. 8.

Jessica Van Brummelen. 2019. *Tools to Create and Democratize Conversational Artificial Intelligence*. Master’s thesis. Massachusetts Institute of Technology. Cambridge, MA.

Jessica Van Brummelen. 2022. Appendix: What Do WEIRD and Non-WEIRD Conversational Agent Users Want and Perceive? Towards Transparent, Trustworthy, Democratized Agents. https://gat.github.io/jessvb/faitd4d759101006d736d196f5d4d7d5d3. Accessed: 2022-09-16.

Jessica Van Brummelen. 2022. *Empowering K-12 Students to Understand and Design Conversational Agents: Concepts, Recommendations and Development Platforms*. Ph. D. Dissertation. Massachusetts Institute of Technology. Cambridge, MA.

Jessica Van Brummelen, Viktória Tabunshchyk, and Tommy Heng. 2021. “Alexa, Can I Program You?”: Student Perceptions of Conversational Artificial Intelligence Before and After Programming Alexa. In *Interaction Design and Children (Athens, Greece)* (IDC ’21). Association for Computing Machinery, New York, NY, USA, 305–313. https://doi.org/10.1145/3459990.3460730

Jessica Van Brummelen, Mingyan Claire Tian, Maura Kelleher, and Ngii Hoang Nguyen. 2022. Learning Affects Trust: Design Recommendations and Concepts for Teaching Children—and Nearly Anyone—about Conversational Agents. https://arxiv.org/abs/2209.05063. In *Human-Computer Interaction*.
[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Advances in neural information processing systems. Advances in neural information processing systems, Long Beach, California, 5998–6008.

[59] Sarah Theres Völkel, Ramona Schödel, Daniel Buschek, Clemens Stachl, Verena Winterhalter, Markus Bühner, and Heinrich Hussmann. 2020. Developing a Personality Model for Speech-Based Conversational Agents Using the Psycholexical Approach. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI ’20). Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376210

[60] John K. Waters. 2014. Idc Study Counts the World’s Developers: 11 Million Pros. https://cacm.acm.org/news/171546-idc-study-counts-the-worlds-developers-11-million-pros/fulltext. Accessed: 2022-09-12.

[61] David Wolber, Harold Abelson, and Mark Friedman. 2015. Democratizing Computing with App Inventor. GetMobile: Mobile Comp. and Comm. 18, 4 (Jan 2015), 53–58. https://doi.org/10.1145/2721914.2721935

[62] Xizhu Xiao, Porismita Borah, and Yan Su. 2021. The dangers of blind trust: Examining the interplay among social media news use, misinformation identification, and news trust on conspiracy beliefs. Public Understanding of Science 30, 8 (2021), 977–992. https://doi.org/10.1177/0963662521998025 arXiv:https://doi.org/10.1177/0963662521998025 PMID: 33663279.

[63] Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 256, 16 pages. https://doi.org/10.1145/3411764.3445445

[64] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. NAACL-HLT (1) 1 (2019), 629–634.