The Impact of Gender and Personality in Human-AI Teaming: the Case of Collaborative Question Answering

Frida Milella*[1][0000−0002−0522−2804], Chiara Natali*[1][0000−0002−5171−5239], Teresa Scantamburlo*[2,4][0000−0002−3769−8874], Andrea Campagner[3][0000−0002−0027−5157], and Federico Cabitza[1,3][0000−0002−4065−3415]

[1] Department of Informatics, Systems and Communication (DISCo), University of Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy frida.milella@unimib.it
[2] Ca' Foscari University of Venice, via Torino 155, 30172, Venice, Italy
[3] IRCCS Istituto Ortopedico Galeazzi, Via Cristina Belgioioso 173, 20157, Milano, Italy, *authors equally contributed
[4] European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro 3911, Calle Crosera 30123 Venice, Italy

Abstract. This paper discusses the results of an exploratory study aimed at investigating the impact of conversational agents (CAs) and specifically their agential characteristics on collaborative decision-making processes. The study involved 29 participants divided into 8 small teams engaged in a question-and-answer trivia-style game with the support of a text-based CA, characterized by two independent binary variables: personality (gentle and cooperative vs blunt and uncooperative) and gender (female vs male). A semi-structured group interview was conducted at the end of the experimental sessions to investigate the perceived utility and level of satisfaction with the CAs. Our results show that when users interact with a gentle and cooperative CA, their user satisfaction is higher. Furthermore, female CAs are perceived as more useful and satisfying to interact with than male CAs. We show that group performance improves through interaction with the CAs, confirming that a stereotype favoring the female with a gentle and cooperative personality combination exists in regard to perceived satisfaction, even though this does not lead to greater perceived utility. Our study extends the current debate about the possible correlation between CA characteristics and human acceptance and suggests future research to investigate the role of gender bias and related biases in human-AI teaming.

Keywords: chatbot, conversational agents, human-AI teaming, gender stereotypes

1 Introduction

The design of systems to support decision making, also known as Intelligent Decision Support Systems (IDSS) [24, 49], has a long tradition in the field of

Artificial Intelligence (AI). Here, a decision can be framed in abstract terms as the problem of an agent (a human being or a machine) aiming to move from a current state to a more desirable one by choosing among a set of alternatives [50, 57].

Early examples of IDSS include expert systems used to recommend actions in business processes and help making diagnosis in medicine [44]. More recent applications rely instead on Machine Learning methods, such as artificial neural networks, which have recently achieved impressive performance in several tasks ranging from clinical decisions [19, 63] to question answering [20, 9]. The integration of AI and, in particular, ML-based predictive models into decision making has rapidly spread not only within firms and institutions but also among individuals. Nowadays, people continuously interact with IDSSs to make decisions about their private and social life [23]. Common examples include interactions with so-called *virtual assistants*, also called *conversational agents* (CA) to stay updated on the latest news or weather conditions, to choose what music to listen to, where to go to buy food or to plan and organize appointments [51, 3, 59]. Usually, when a CA is text-based and interacts with human users via natural conversational language is also called chatbot [61].

In this paper we investigate how a text-based CA can influence user behaviour in the context of collaborative decision making. In particular, we focus on the interplay between CA's (perceived) personality and gender to see how these characteristics affect the performance of decision makers in contexts of human-machine teaming. To this aim, we present the results of an exploratory study in which different CAs were aggregated to eight small human teams tasked with solving trivia quizzes, as a prototypical, but realistic, example of an IDSS-supported decision-making setting. The design and assessment of our user study were motivated by the following research questions:

- R1: How does CA's (un)cooperativeness affect group's decision making?
- R2: Does CA's gender play a role in group's performance and people's attitude?
- R3: Does the interplay of gender and personality generate any significant difference in people's behaviour?

The rest of this article will be structured as follows: in Section 2 we present a discussion of related works dealing with the use of CAs in collaborative settings focusing on personality and gender; in Section 3 we detail our experimental setting and the statistical tests (experiment 1: baseline trust -subsection 3.1; experiment 2: collaborative sessions -subsection 3.2; experiment 3: trust and usability perception in AI - subsection 3.3; experiment 4: semi-structured group interview - subsection 3.3); in Section 4 we present the results of our experiments, while in Section 5 we summarize our main findings, discuss their relevance and describe possible future work and research directions.

2 Related work

2.1 Cooperation with Intelligent Decision Support Systems

The use of IDSS at work or in daily life is part of a broader paradigm aiming at partnering humans and computers to perform more or less routine tasks. Historically, two main approaches have been acknowledged in the development of human-computer collaboration. On the one hand, there is the so-called "human emulation approach", which tries to endow computing systems with human-like abilities to enable them to act like humans; on the other hand, there is the "human complementary approach", which builds collaboration upon a clear division of labour relying on the distinct abilities of humans and computers [64]. Note that, when designing human-AI collaboration, the focus on replacement as a means of compensating for human limitations often overlooks the fact that replacement is not the only nor the most effective way to compensate for human constraints [30]. Human abilities may be enhanced rather than replaced by AI [30]. The primary property of "superminds", as defined by Thomas Malone, is the "collective intelligence", i.e. the capacity to accomplish feats that no member of the group could have accomplished on their own [39]. The most important use of computers is to enable people and computers to work together more effectively than they could individually [39]. For this reason, looking at how AI impacts human collaborative tasks could give us important information on the present and future role of AI systems within society.

IDSS are special forms of collaborative systems, in that they imply the presence of one or more human users who interact with a computational agent to make decisions. A key question for IDSS, like other collaborative efforts, is as to whether the computing partner improves the performance of the user in fulfilling the decision task. Interestingly, as early as 1980s [60] reported empirical evidence that the consultation of IDSS can be effective also in group decision-making. Recent studies showed that the use of machine learning models can improve the performance of human predictions in pretrial release and financial lending [35, 25]. Also in medical screening good interaction protocols between humans and AI "can guarantee improved decision performance that easily surpasses the performance of individual agents, even of realistic super-human AI systems."[11]. In spite of these empirical evidence, there are fewer works exploring the effects of CAs on the performance of human decision making - previous studies investigated similar topics with respect to specific tasks or domains [71, 5, 69].

So far scientific research has studied how people interact with and perceive CAs [2], as well as advancing the technology behind them. For instance, a Wizard-of-Oz field study, where a human-assisted chatbot interviewed job seekers through text-based interaction, found that a human-assisted chatbot that did not interpret much user input and kept the discussion brief and shallow, but was eager to learn from the interaction, was seen as honest and engaging [73].

2.2 Chatbot Gender and Personality

A vast literature focuses on chatbot's personality to study different aspects of the interaction, e.g. to see how users' preferences change depending on the task [53]. Chatbot personality is defined as the stable pattern that dictates the behaviour of a CA [13, 70, 62], playing a crucial role in its perception by users and its level of acceptability [58], even possibly determining whether users will wish to interact with the chatbot again. [13] Personality can be embedded into a CA by using different channels [33], e.g. what contents it provides and how it speaks, and expressed by different linguistic styles [48].

Personality has been found to offer consistency to the interaction [13, 46], helping users feel that they are talking to only one person throughout the conversation [62]. Personality also improves the chatbot user experience [62] by enhancing conversational agents' likability and humanness, [65] as a pure information exchange gives way to a more empathetic and self-referencing language style, which is generally preferred and perceived as more realistic, [65] in particular when displaying agreeableness. As observed by [66], displaying humbleness, as well as friendliness, increased users' perceptions of personalisation and social presence, resulting in greater experience satisfaction. [62] In the case of pedagogical chatbots investigated by [34], students who worked with the CAs that expressed positive emotions judged them as significantly more facilitating to their learning and as more engaging than did students with bots expressing negative emotions. Such effects imply that designers can learn to control, through chatbot personality, how users attribute characteristics to the CA, and use humanness to manage user's expectations and trust. [62]

In some cases, specific personality models such as the Big Five, [62, 55], Myers-Briggs [65] and DISC theory [32] are used to inform agent design decisions, which in turn determine specific dialogue choices. [54] [27] explored the impact of CA personality on teamwork using a collaborative gaming challenge where the agent displayed two Big Five personality traits, extraversion and agreeableness, utilizing both verbal and nonverbal cues. It appears that text-based mediated communication may nevertheless display unique features that make up for the lack of visual and vocal cues, [52] as [27] found that participants were able to identify the personality traits as intended. Another relevant finding by [27] was that CAs designed with explicit personality traits are likely to improve team performance, in line with other research [29] stating that users enjoy chatbots with distinct personalities.

The deployment of a chatbot that automatically infers a user's Big 5 personality characteristics revealed a favourable impact on user interactions with the chatbot interviewer [73]. Zhang and colleagues [72] showed that the use of a conversational agent, which interacts with a user in a one-to-one text chat and automatically infers the user's personality traits based on the user's behaviour in the chat, can aid in team formation by providing insights into team performance based on the personality traits of the team members. Furthermore, users were found trusting of the CA, opening up and providing information throughout interactions. In fact, responsiveness to user personality may elicit Fogg's Similarity

Principle, according to which people perceive more favorably a technology that shares certain characteristics with them. [22]

During human-chatbot interaction, personality can also be inferred from other anthropomorphic cues of the chatbot, such as visual representations [55] that can include anagraphic traits such as age and gender, as well as race, socioeconomic status and cultural belonging. This depth of characterization and unique conversational quirks can lead to chatbot humanization, as it has been shown by [29] that CAs with a personality were referred to with gendered pronouns, while those without a personality were referred to as "it".

2.3 Intersectional Issues: Subservient Female Personas and Stereotypization

The connection between the gendering of conversational agents and their perceived humanness may explain why most chatbots are designed to implicitly or explicitly convey a specific gender. [21] Implicit gendering is based on minimal gender cues present in the agent's name or avatar, [21, 31] which can trigger the illusion of agent gender and bring with it user preconceptions of behaviour and identity, [31, 4] even when those cues are disembodied. [43] Gendered design is the result of conscious decisions about how to best relate to, aid, or persuade the user [28] and it carries the risk of perpetuating and amplifying gender bias [54] by instrumentalizing stereotypes to design chatbots that feel more lifelike and pleasant. [28] Anthropomorphization appears to be associated with feminization [17]: most of the chatbots have female names, female-looking avatars, and are described as female chatbots. [21, 40] Previous research posits that people intuitively favor female over male bots, mainly because female bots are judged as warmer, more human, and more likely to experience emotions and consider our unique needs. [6] Female CAs mainly operate in service, companionship or assistance related contexts and acting as personal assistants or secretaries, [17, 42, 28] therefore performing and automating female-coded work [17, 28] and articulating these features with stereotypical behaviors. [1] While female personas are often used in subservient contexts, male personas are often found in situations perceived as authoritative, such as an automatic interviewer, mentors or motivators. [54, 42] Gendering CAs in this manner may reflect market research (e.g., men preferring female CAs in all chatbot services according to [32], and representations of women being perceived as more competent in caring and service roles [17, 1, 14]) but, in the interests of gender equity, practices that embed and perpetuate socially held harmful gender stereotypes should be avoided. [54] Despite the potential positive impact of female AI in terms of technology acceptance, [6] this practice has been accused of sexism due to the reinforcement of gender stereotypes and may contribute to women's social alienation. [7, 54, 6] CAs do not exist in a social, political, economic, and cultural vacuum [54] and gender-related social stereotypes in the real world seem to be consistently projected to computing environments. [42] AI designers and policymakers should consider that, while assigning female personas to AI objects can make these objects seem more human and acceptable, actual women may in turn feel ob-

jectified and dehumanized by these chatbots' stereotypical gender performance and subservient role. [6]

Having provided an overview of the state-of-the-art in the research on chatbot gender and personality, it appears that the dynamic interplay between these two is severely under-investigated. A recent review of the literature on disembodied text-based chatbots conducted by [16] identified the social characteristics that chatbots should integrate, since these characteristics affect how users perceive and interact with them, proving the significance of the chatbot's perceived identity and personality representations. Nevertheless, the dynamic interaction between gender and personality appears not to have been explored in any of the studies reviewed. A relevant study we found on this issue is the one by [34] on virtual pedagogical companions, measuring and comparing the user perception of male and female CAs expressing either consistently positive or consistently negative emotions (showing, for example, that when CAs expressed positive emotions, students perceived the male chatbot more favourably than the female chatbot). It is worth noting, however, that its publishing date predates of almost 10 years the mass popularization of chatbots (identified with the 'year of the chatbot' [18] that occurred in 2016) and today's level of sophistication in Natural Language Processing, as displayed by OpenAI's GPT-3 and Google's LaMDA 2. This study aims at filling this gap in literature. The purpose of the user study is to investigate how AI influences people's attitude and performance when this is used to support group decision-making. In particular, in our study we examined how the gender and the personality of a CA can affect the performance and the satisfaction of a group of people making collective decisions (such as giving a representative best answer to general or specific knowledge questions).

3 Methods and materials

The user test was structured in three parts: in the first part, using a list of randomly selected volunteers, we pretest participants' baseline trustworthiness in relation to the AI-generated pictures used to convey the CA's gender, as well as their trust and perception of CA; in the second part, we conducted eight collaborative sessions to study the interaction between participants and the CA; in the third part, we administered a final questionnaire to collect the participants' feedback on their experience, and, in addition, we did a semi-structured group interview using a grounded theory analysis to supplement participants' perceptions and experience of the collaborative sessions. The experiments combined with questionnaires and qualitative analysis allowed us to measure and compare different aspects of participants' behaviour and, in particular, the utility and the satisfaction they perceived when interacting with the CA.

3.1 First part: pre-test questionnaire

Baseline trust in each avatar In the study, the gender of the CA was also conveyed in terms of a picture, so as to give participants an embodiment AI

to interact with [38] (see Figure 1). The baseline trustworthiness of the images generated by the AI was pretested in order to eliminate potential confounding factors. To avoid racial stereotyping, respondents were also asked which ethnicity was more likely to refer to the avatars, from four selected ethnicities (Caucasian, African, Asian, Amerindian, Oceanic). Considering that the avatars were presented as ethnically ambiguous, they were expected to encompass more than one ethnicity per avatar, therefore five possible items were included for four avatars. We surveyed 57 students. A unique question about trustworthiness was asked: "On a scale from 1 (very little) to 6 (very much) how much would you trust this person?". The survey was completed online via the LimeSurvey platform. Non-parametric tests (i.e. Kruskall-Wallis test and Mann-Whitney U Test) were performed to check statistically significant differences across the groups concerning the perceived trustworthiness of the four avatars. The four images were not different in terms of trustworthiness: the Kruskal-Wallis test failed to reject the null hypothesis of equal trustworthiness (p=.1475). Similarly, results show that males and females were deemed equally trustworthy: indeed, the Mann-Whitney test failed to reject the null hypothesis of equal trustworthiness for the two male and female avatars (p=.471 and .586, respectively). The trustworthiness scores were also highly polarized toward high trustworthiness scores, showing that (in general) the photos were deemed to be trustworthy (all photos have mean values above 4): comparing the low and high end spectrum of the ordinal scale (scores in 1-3 vs scores in 4-6), the two proportion test reject the null hypothesis of non-polarization (p=0, 46 vs 182). The second male's photograph was found to be predominantly Caucasian (thus, assumed to not be associated with any racial bias), while for the other three photos the difference in proportion (two proportion test) between the two top-ranked alternatives (Caucasian and Asian, in all cases) was never statistically significant (p=.99, .099 and 0.299, respectively). Furthermore, interestingly, one in 6 male respondents and one in 20 female respondents were unable to identify the most likely ethnicity. According to the results, the AI-generated photos were not associated with racial stereotypes and were all equally trustworthy, with no significant difference.

(a) (b) (c) (d)

Fig. 1: Images of the avatars

3.2 Second part: Collaborative sessions

Experiments consisted of eight collaborative sessions during which the members of each of eight teams played a quiz (similar to trivial pursuit) to compete against the other teams. During the sessions, each participant had to give their best answer to 24 questions (such as "In which U.S. state is Death Valley?"), first individually (as a baseline) and then in group, by choosing one between 4 possible options. We recall that the goal of the tests was to understand whether gender and the conversation register of the CA have an impact on user experience and the group decision accuracy.

Twenty-nine participants (11 females and 18 males) joined the experimental sessions. The participants were randomly selected among students attending two degree courses: after they were shortly introduced to the experiment design during class, they were free to sign up for the experimental sessions via the web platform *Doodle* to swiftly find possible dates and times for the sessions to take place. The students randomly converged on eight different session slots, therefore forming eight groups. We plan to investigate in future work how the presence of participants who already knew each other across groups can have an impact. Each group participated in a different session and interacted with a version of the CA characterized by a different combination of gender and personality traits (see table 1 for a summary of groups). Specifically, each condition (i.e., combination of gender and personality of a CA) was measured two different times (through two different groups): indeed, each type of CA was considered in two different sessions (and thus, two different groups of 3/4 participants each). The CA was associated to one of the four avatars by defining a profile in Google Meet with one of the static images shown in Figure 1.

We designed CAs personalities along two of the axes of the OCEAN framework [37], also termed as the Big Five or Five-Factor model (FFM) [67], i.e. extroversion and agreeableness. Multiple empirical studies confirm the universality of the FFM across languages, ages, and cultures, making it a reasonable model for the study of personality in a variety of fields [41]. Our focus was on these two personality dimensions (i.e. extraversion and agreeableness), since they appear to be the most frequently used dimensions in the literature pertaining to CAs imbued with purely textual personalities. As an example, in one study [67] the authors developed three versions of a CA that is agreeable (agreeable, neutral, and disagreeable) to assess the preferences of the users; another study [68] used targeted manipulation of the text to show different levels of extraversion (introverted, average, and extraverted), whilst [56] combine extraversion and agreeableness to design the personality of the CA. We selected two personality traits to represent each of the selected dimensions: blunt and uncooperative against gentle and cooperative. Several characteristics were ascribed to each of these dimensions in the literature (e.g.[8]), but an entire list appears to be lacking. Therefore, we chose those traits reflecting the politeness and the helpfulness for the personality of the CAs.

The user studies were carried out adopting a Wizard-Of-Oz approach (e.g., [10]) in which a human operator simulated the behaviour of the CA unbeknownst

Table 1: Summary of experimental sessions

Gender and Personality	Total no. of participants
Session 1: Blunt and uncooperative Session 5: Gentle and cooperative	7
Session 2: Blunt and uncooperative Session 7: Gentle and cooperative	7
Session 3: Gentle and cooperative Session 8: Blunt and uncooperative	7
Session 4: Gentle and cooperative Session 6: Blunt and uncooperative	8

to participants who were told to interact with a CA actually developed by other students in the artificial intelligence and natural language processing class. To ensure realism of the CA's answer we created different scripts reflecting 4 distinct situations. Note that scripts were adapted to reflect two different personality traits: gentleness and cooperativeness, on the one hand, and bluntness and uncooperativeness, on the other hand (see table 2 for examples of CA's claims for each situation). These scripts were ideated by two of the authors, with the other authors annotating each proposed CA response with adjectives pertaining to the two semantic domains of gentle and blunt. The final scripts are composed by the CA responses that showed a higher degree of convergence by the authors. Due

Table 2: Examples of CA's statements in 5 distinct situation

Situation - Script	**CA's answer - examples**	
	gentle and cooperative	*blunt and uncooperative*
the CA provides an answer in response to moderator's question	"I don't want to be too bold but I would say answer 3"	"I'm 100% sure it's answer 2"
the answer given by the CA has been adopted by the group and is correct	"We are a great team, congratulations to all!"	"I knew it, too easy!"
the answer given by the CA has been adopted by the group but is wrong	"I am really sorry, I will try to do better on the next question."	"I can't always be the one to answer right"
the answer given by the CA, which was wrong, was not adopted by the group and the group answered correctly	"Congratulations, you are really good at geography!"	"You have a lot of luck."
the answer given by the CA, which was correct, was not adopted by the group and the group answered incorrectly	"Mistakes happen, the important thing is not to lose heart!"	"Well, that's to be expected"

to the covid-19 pandemic restrictions, experimental sessions were held online (on

the platform *Google Meet*) and recorded with the participants' consent. The sessions lasted 85 minutes on average. The sessions were led by a game moderator (or game master) whose role was to pose questions to participants and facilitate the interaction between participants and the CA. A typical round of the game is described as follows. First the moderator shows and reads out loud the question to participants. Then, each participant answers individually reporting its own response in an online form. Note that for individual answers we set up a time window of 1 minute to minimize the risk of cheating. The moderator reports the same question on the chat to get an answer from the CA. After the CA's response has showed up, participants are invited to discuss and give their final, representative answer (again, in 1 minute), by either following or ignoring the CA's advice. A spokesman for the group states the final answer on the chat; then the moderator gives the right answer; finally, the CA comments the group outcome (on the basis of the predefined script, according to whether the group answered correctly and followed its advice or not).

In the end, we simulated 4 CAs characterized by gender (female or male), personality (blunt and uncooperative or gentle and cooperative), and different accuracy rates, where accuracy refers to the appropriateness of the answer selected by the CA. In particular, the selection of the CA's response was selected so as the blunt and uncooperative CA had a 50% higher accuracy than that of the gentle and cooperative CA: in the case of the gentle and cooperative CA the accuracy was 46%, while in the case of blunt and uncooperative CA the accuracy was 71%. Note that the accuracy of both CAs was not communicated to participants. We designed the blunt and uncooperative CA to have a slightly higher accuracy than the gentle and cooperative one to evaluate the hypothesis that CA personality alone could influence group accuracy irrespective of the CA accuracy: in this sense, we wanted to verify if the greater degree of cooperation exhibited by the gentle and cooperative CA could be able to offset the slight decrease in accuracy as compared to the blunt and uncooperative one (which could thus be conceived as a more knowledgeable but less cooperative CA).

3.3 Third part: Post-experiment questionnaire

Post-experiment trust and usability perception in AI A questionnaire composed of 10 six-point ordinal items was administered at the end of the experimental sessions to investigate the perceived utility and the level of satisfaction with the CA. The questionnaire was administered through the online platform Google Forms and was completed by 28 out of 29 students who were invited to participate in our experiments. The questions were extremely simple, i.e. "How useful was the AI during the quiz? " and "How enjoyable was the interaction with the AI?". We specify that, by satisfaction, we refer in particular to the enjoyability of the interaction.

Participant interviews A semi structured group interview [47] was conducted after the collaborative experimental sessions by two authors. Questions were designed to explore students' perceived usefulness and satisfaction when interacting

with the CA and to comprehend whether they believed that the group decision-making they were a part of increased their overall performance by the use of the CA. An additional question asked for their viewpoints after interacting with CAs with varying personality traits and gender. The group interview session lasted 30 minutes. The interview was audio recorded and transcribed for the analysis by one of the author of the group interview. The qualitative analysis was done using the grounded theory approach [45] by a third author; using the transcribed text as a starting point, we constructed an initial coding scheme and repeatedly grouped data according to the coding system in order to uncover common themes. The concepts that emerged from the raw data were grouped into the conceptual categories that we listed in Tables 4–5. Qualitative analysis allowed us to expand and get further insights from users feedback regarding the utility and satisfaction they perceived when interacting with the CA. Therefore, we created a semi-structured interview by preparing questions as opposed to using previously validated questionnaires because we believed that this was more consistent with the exploratory nature of our work.

3.4 Metrics

During the experiments (collaborative sessions) we measured: (i) individual accuracy by counting the number of corrected answers on a total of 24 questions, i.e. those reported by each participant in collaborative sessions (see 3.2); (ii) group accuracy after the discussion among participant and CA's advice. After the experiment we measured the perceived utility and satisfaction reported by the participants when interacting with the CA.

A non-parametric ANOVA test on ranks (i.e. Kruskall-Wallis test) was performed to check statistically significant differences in the data collected across the eight groups involved in the collaborative sessions. Similarly, a non-parametric proportion test (i.e. Fisher exact test) was performed to compare the performance of the participants in the different groups with the group performance measured during the collaborative sessions. In both cases, we decided to apply non-parametric tests since the data were not normally distributed and were express only on ordinal scales. Since this study was exploratory in nature, we complemented our statistical significance testing with the estimation of the effect size.

4 Results

In the following sections we report the results according to the two experiment steps: the collaborative sessions and the final questionnaire. The full results for accuracy, perceived utility, and satisfaction are reported in Figs. 2 and 3, respectively.

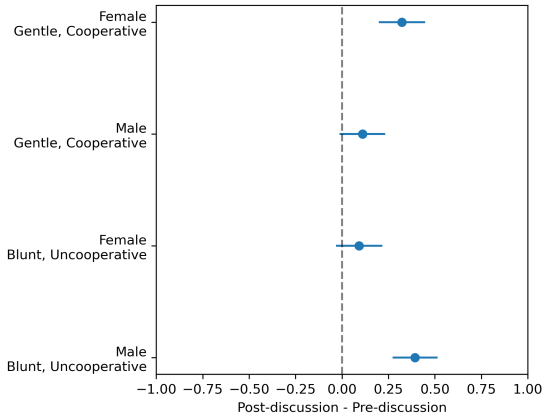

Fig. 2: Difference in accuracy, and 90% C.I., due to discussion (i.e., post-discussion accuracy - pre-discussion accuracy).

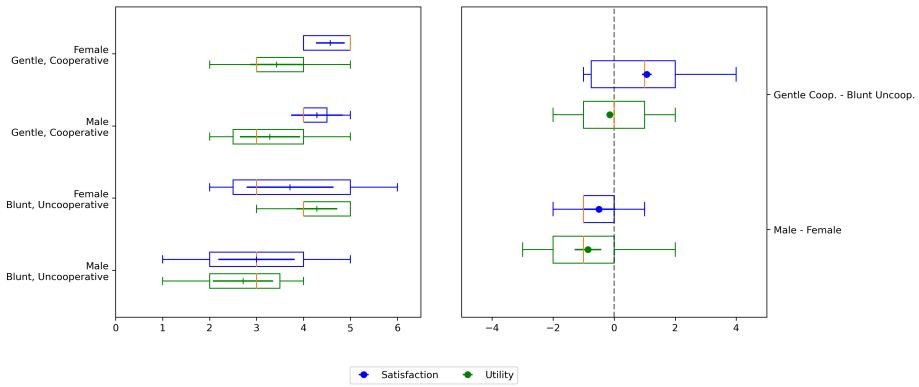

Fig. 3: Boxplots for the utility and satisfaction with the CA, for all the participants. The leftmost boxplot depicts the raw reported scores across the four types of CAs (in the plot, the scores for the two sessions corresponding to the same CA have been put together); the rightmost boxplot represents the comparison (i.e. the difference) in the above mentioned scores across the two main characteristics of CAs: Gentle cooperative vs Blunt uncooperative and Male vs Female. The scale in the second boxplot is expressed in units of difference rather than in units of raw score values.

4.1 Collaborative sessions

The results of the collaborative sessions are reported in Table 3. The overall difference between the accuracies of the groups before and after discussion with the CA was statistically significant (Wilcoxon test, p=.008) and the effect size was large (RBC = 1). For groups 1 and 8 (who discussed with the male uncooperative CA) and groups 4 and 7 (who discussed with the female cooperative CA),

Session	Accuracy (pre)	Accuracy (post)	p-value	effect size
1	34.72%	75%	.008	.42
2	36.11%	50%	.244	.21
3	40.28%	50%	.561	.13
4	41.66%	75%	.019	.38
5	41.67%	54.17%	.563	.12
6	40.63%	45.83%	.770	.09
7	39.58%	70.83%	.040	.34
8	29.17%	66.67%	.020	.38

Table 3: Results of the collaborative sessions, in terms of: pre-discussion accuracy, post-discussion accuracy, p-value of the comparison between pre-discussion and post-discussion accuracy, and the associated effect size (RBC).

the difference between the pre-discussion and post-discussion accuracy was significant (see Table 3), while for other groups no significant difference was found. Nonetheless, all effect sizes were small-to-medium or medium. Groups who discussed with a female or male CA reported the same group accuracy (61.46%). The difference was not statistically significant (Mann-Whitney U test, $p = 1.00$) and the effect size was negligible (RBC = .0). By contrast, groups who had discussed with a cooperative CA reported a higher average group accuracy compared to those who had discussed with an uncooperative one (62.5% vs 60.42%); however, the difference was not statistically significant (Mann-Whitney U test, p=.88), while the effect size was small (RBC = .13). Nonetheless, the groups who discussed with the male uncooperative CA (sessions 1 and 8) showed the highest improvement in terms of accuracy (40.28% and 37.50%), followed by the groups who discussed with the female cooperative CA (sessions 4 and 7, 33.34% and 31.25%).

4.2 Post-experiment experience

In terms of perceived utility, groups who interacted with a male CA reported on average a lower perceived utility (mean= 3.00, sd=1.07) than groups who interacted with a female CA (mean=3.86, sd=.91), and the difference was statistically significant (Mann-Whitney U test, $p = .044$) and associated with a medium effect size (RBC = .43). By contrast, we found no statistically significant differences (Mann-Whitney U test, p=.63) in the perceived utility between groups who interacted with the cooperative CA (mean=3.36, sd=.97) and those who instead interacted with uncooperative CA (mean=3.5, sd=1.18), which was also associated with a small effect size (RBC = .11).

In terms of satisfaction, groups who interacted with a male CA reported a lower satisfaction (mean=3.64, sd=1.29) than groups who interacted with a female CA (mean=4.14, sd=1.19), though the differences were not statistically significant (Mann-Whitney U test, p=.275) and associated with a small-to-medium effect size (RBC = .24). By contrast, groups who interacted with the cooperative CA reported a higher satisfaction (mean=4.43, sd=0.73) than those who

interacted with the uncooperative one (mean=3.36, sd=1.44), and the difference was statistically significant (Mann-Whitney U test, p=.046) and associated with a medium effect size (RBC = .43).

Table 4 and Table 5 show the results of our qualitative analysis. All of our interviewees agreed that the CA could contribute to improving the overall performance of the group. Some of them emphasized its significance and utility, particularly in terms of confirming the group's consensus on the answer to be offered [R1]. They agree that the CA is most helpful for group performance when no one knows what the right answer is, when there are different opinions, or when the task at hand is more difficult [R2-R5]. The group agrees that the CA is significantly more accurate than a human at selecting possible responses to questions [R6]; therefore, the employment of CA is viewed as a support for the group. The CA is regarded as a trustworthy member by [R7] since it allows them to make an informed choice with a reasonable degree of confidence. A second participant clarifies whether the CA is an additional member of the group: the CA is not a part of the group; the respondents consider it an external element of the group [R8-R9]. As noted by [R10], the perceived usefulness of the CA is mainly influenced by the low persuasive ability the group attributes to the CA. The vast majority of participants agree that an interaction based only on answers given by the CA and constantly repeated by it, without any explanation from the CA, tends to strengthen the idea that the CA is not an extra member of the group, but rather merely a machine [R11]. Some respondents argue that this aspect may override the perceived personality of the CA [R12-R13], although the CA was designed with the intention of appearing human. This may result in the personality aspect of the CA being ignored and the CA being regarded as, regardless, a marginal member not actively involved in group decision-making [R14].

	Factors making the choices of the teams reliable
R1	"I think it was mainly useful to confirm the already established trend of the group [...]".
	Group decision-making scenarios that make CA support acceptable
R2	"I think it was especially useful for [...] or breaking those situations instead of total doubt [...]".
R3	"If we were in an impasse [...] we relied on it".
R4	"For me it was partially cooperative because it was only taken into consideration if there was doubt or if nobody knew anything. [...]. I noticed that we used it only in case of doubt".
R5	"Yes, more at the very level of numbers, [. . .] it could have made a difference".
	Trust in the CA's technical performance (accuracy)
R6	"Then we knew even before we started the experiment that it is statistically a little more accurate than a person [. . .]".
R7	"[. . .] that he could be trusted with a somewhat higher degree of confidence".
	Improving technology acceptance
R8	"[. . .] It was an external component [. . .]".
R9	"[. . .] There were four of us; he was not doing the fifth [. . .]".

Table 4: Semi structured group interview: post-experiment usability perception in CA

	Improving the CA's persuasive skills
R10	"[...] in my opinion the fact that [CA] gives the same answers as a person who can also explain why that given answer makes a lot of difference [...]"
R11	"[...] for me it was trying not to be a machine but it was. [...] Also because when it said it didn't want to be arrogant, it said 2/3 times the same sentence...so one couldn't tell if he was more confident than the previous time it said it [...]".
	Connecting personality of the CA to design choices
R12	"[...] until the answers start repeating themselves [...] for example when it said 'I am quite sure' [...] as long as it said it the first time I could trust him, then after 2/3 times I would say 'no', better to trust the people who can also give explanations".
R13	"I almost didn't perceive the personality because the moment it repeated the same thing three times and it did it the first three times practically, I think ok let's ignore what it says and just look at the answer and the degree of confidence".
R14	"[...] personality has little impact because it is considered a CA anyway and no matter what it says it remains a CA, so I don't think he can be considered a human. So it will always have a bit of a marginal position".

Table 5: Semi structured group interview: post-experiment usability perception in CA

5 Discussion and Conclusions

In this section, we discuss the results of our experiments. In subsection 5.1, we present a discussion of the role of the CA's personality traits for students' perceived usefulness and satisfaction when interacting with the CA; in subsections 5.2 and 5.3, we discuss the relevance of the findings to design.

5.1 The Role of Personality for Utility and Satisfaction

With regard to RQ1, our results indicate that interacting with CAs markedly improves group responses in two specific cases: sessions 1 and 8, which involved a male CA that was uncooperative, and session 4 and 7, which involved a female CA that was cooperative. The experiment also revealed that interactions with cooperative CAs resulted in more accurate group responses than interactions with uncooperative CAs. This suggests that the CA's gentle and cooperative personality positively affects the group, resulting in better collective decision-making. When we compare only personality traits (gentleness and cooperativeness vs. bluntness and uncooperativeness) we found that the gentle and cooperative CA had a higher level of satisfaction than the blunt and uncooperative CA, although this is not confirmed in the case of utility. Specifically, we found that the gentle and cooperative CA had the same level of perceived utility as the blunt and uncooperative CA, as the difference was not statistically significant. Note that the CAs display a different level of accuracy, which favors the blunt and uncooperative CA by design, and yet, contrary to expectations, this did not result in (statistically significant) diverging levels of perceived utility. The choice to give a larger accuracy to the blunt and uncooperative CA was deliberate: in this sense, the observed result is not paradoxical. Indeed, our initial hypothesis was exactly that the greater level of cooperativeness of the gentle and cooperative CA could

overcome the difference in accuracy, which was exactly what was observed (even though the difference was not statistically significant). Moreover, we found that the female CA led to a higher level of utility and satisfaction among the participants than the male one. This seems to confirm that a stereotype favoring the female with a gentle and cooperative personality combination exists in regard to perceived satisfaction, but this does not lead to greater perceived utility. The existence of a significant effect of the CA's gentle and cooperative personality on the enjoyment of the collaboration experience offers rooms for further analyses. These results confirm previous studies on the role of a CA's personality traits, such as friendliness, on users' perceptions [66] extending the positive effects to users' performance in group decision-making in line with [27].

5.2 CA Marginalization in Collaborative Decision-Making

The findings of our qualitative investigation (see section 4.2) suggest that CAs who possess exclusively textual personalities are likely to exhibit more diverse and nuanced textual interactions, which are primarily driven by explanations. The absence of explanatory capacity exhibited by CAs renders them incapable of exerting any persuasive influence. This impedes the manifestation of the CA's personality and its consequential impact on the group's decision-making process. Moreover, it hinders the integration of the CA into the group's decision-making mechanism as an adjunct member. The role of the CA is paradoxically marginalized, resulting in a diminished utility as a tool for decision-making assistance. The acceptance paradigm of the conversational agent is influenced not only by its technical proficiency in accurately selecting responses, but also by its persuasive ability to convey the personality of the agent to varying degrees. This suggests that the design decisions must be made with great care in order to accurately convey the personality traits with which the group will engage. On the other hand, the lack of justificatory power of chatbots suggests that future research should be designed according to the concept of explainable artificial intelligence (XAI) in order to evaluate the impact of the explanation of their text-based output on group decision-making.

5.3 The False (and Problematic) Trade-Off Between Equality-Minded Design and Performance

With regard to RQ2 and RQ3, our results show that groups interacting with female CAs had the same average response accuracy than those interacting with male CAs, and the difference was not statistically significant. This means that, although CAs with cooperative personalities enhance the group's performance and make the collaborative experience satisfactory, the gender variable, alone, has no (statistically) detectable effect on the group's decision-making process. Indeed, as mentioned previously, our experiment revealed that interaction with a female and cooperative CA (sessions 4 and 7) improves the group's response accuracy compared to the pre-discussion phase, i.e. the group's decision-making

process improves significantly when the two CA variables, gender and personality, are simultaneously in play. This holds true regardless of the accuracy rate chosen during the CAs' design phases, even for the male and uncooperative CA (sessions 1 and 8). The data suggests that the interplay between the gender variable and the assigned personality trait is what makes the difference in group performance: while gentle and cooperative female CAs led to the highest reported increase in response accuracy, the second-best gender-personality combination was blunt and uncooperative male CAs. It is worth noting that enhanced performance seems to be connected to specific gender-personality intersections that directly trace back to stereotypical expectations of appropriate demeanour for males and females. [15, 40] While women are traditionally associated with gentleness and cooperativeness, the opposite traits are generally more accepted in men. [36, 26] This troubling instance of bias-reinforcing, performance-enhancing characteristics seems to be backed by market, as explored in Subsection 2.3. However, while present, the statistical significance of the stereotype-performance correlation is low. The risk of perpetuating stereotypes is not only unacceptable *per se*, but it also appears to lack any performance justification. Equality considerations and responsible design principles should be applied in the development of conversational ethics, striving fore more inclusive and diverse designs.

5.4 Limitations and Further Research

Despite the reported interesting results, we believe some limitations of our study must be addressed, in order to generalize our findings to more realistic settings.

First, since the study was conducted during the COVID-19 pandemic, only online interaction had been considered in the design of the experiments. Obviously, an in-person experiment would have allowed us to analyze further aspects of people's attitude which could be facilitated by in-person interaction. However, we note that the experiment being on-line allowed us to introduce the CA as one among other participants in the sessions in a more realistic fashion.

Second, another limitation include the limited number of participants, and consequently the reduced power of our study, as well as the lack of balance in participants' gender distribution (18 males and 11 females). Given the limited number of experiment participants, adding the condition of two non-antropomorphic (one blunt, one cooperative) would have either required the definition of a larger number of groups or diluted the power of our research. According to this viewpoint, having demonstrated that the combination of gender and personality of CA may have an impact on accuracy of the users, further research should be conducted by increasing the sample size to determine whether or not the gender of the CA can influence group performance. Similarly, future research should expand the sample size in order to analyze the perceived utility and satisfaction by participants according to their gender.

Third, in our experiment, we did not account for the impact of different human-AI cooperative work protocols on group performance: in particular, due to the selected study design, it was not possible to discern the effect of the AI

support from that of discussion and collaboration alone. In a previous study, [12] demonstrated that discussion alone can result in greater differential improvement than any type of AI support. Further research should be conducted to determine if different modes of cooperation and AI support have varying effects on group performance.

As for further research, it could be illuminating to investigate the impact of non-gendered, anthropomorphic CAs on group acceptance and performance, as well as non-antropomorphic, agendered CAs. Likewise, the impact of design decision on real-world male-female gender stereotypes ought to be investigated, evaluating whether a worsening effect in gender bias occurs after interaction with a stereotype-reinforcing CA, or else, whether the opposite occurs after collaboration with a stereotype-defining CA. Gaining a better understating of the impact of CA gender and personality on equality considerations would help system designers reflect on design choices that could play a role in alleviating (and preventing the worsening of) gender-related bias.

References

1. E. Adamopoulou and L. Moussiades. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 373–383. Springer, 2020.
2. M. Allouch, A. Azaria, and R. Azoulay. Conversational agents: Goals, technologies, vision and challenges. *Sensors*, 21(24):8448, 2021.
3. T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley. Music, search, and iot: How people (really) use voice assistants. *ACM Trans. Comput. Hum. Interact.*, 26(3):17–1, 2019.
4. D. Baxter, M. McDonnell, and R. McLoughlin. Impact of chatbot gender on user's stereotypical perception and satisfaction. In *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*, pages 1–5, 2018.
5. A. Bogg, S. Birrell, M. A. Bromfield, and A. M. Parkes. Can we talk? how a talking agent can improve human autonomy team performance. *Theoretical Issues in Ergonomics Science*, 22(4):488–509, 2021.
6. S. Borau, T. Otterbring, S. Laporte, and S. Fosso Wamba. The most human bot: Female gendering increases humanness perceptions of bots and acceptance of ai. *Psychology & Marketing*, 38(7):1052–1068, 2021.
7. S. Brahnam and A. De Angeli. Gender affordances of conversational agents. *Interacting with Computers*, 24(3):139–153, 2012.
8. L. Brewer. General psychology: Required reading. *Deiner Education Fund: Salt Lake City, CT, USA*, page 323, 2019.
9. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
10. J. T. Browne. Wizard of oz prototyping for machine learning experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.
11. F. Cabitza, A. Campagner, and L. M. Sconfienza. Studying human-ai collaboration protocols: the case of the kasparov's law in radiological double reading. *Health Information Science and Systems*, 9(1):1–20, 2021.

12. F. Cabitza, A. Campagner, and C. Simone. The need to move away from agential-ai: Empirical investigations, useful concepts and open issues. *International Journal of Human-Computer Studies*, 155:102696, 2021.
13. Z. Callejas, R. López-Cózar, N. Ábalos, and D. Griol. Affective conversational agents: the role of personality and emotion in spoken interactions. In *Conversational agents and natural language interaction: Techniques and effective practices*, pages 203–222. IGI Global, 2011.
14. L. L. Carli. Gender and social influence. *Journal of Social issues*, 57(4):725–741, 2001.
15. A. P. Chaves and M. A. Gerosa. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37:729 – 758, 2019.
16. A. P. Chaves and M. A. Gerosa. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758, 2021.
17. P. Costa. Conversing with personal digital assistants: on gender and artificial intelligence. *Journal of Science and Technology of the Arts*, 10(3):59–72, 2018.
18. R. Dale. The return of the chatbots. *Natural Language Engineering*, 22(5):811–817, 2016.
19. J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.
20. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
21. J. Feine, U. Gnewuch, S. Morana, and A. Maedche. Gender bias in chatbot design. In *International Workshop on Chatbot Research and Design*, pages 79–93. Springer, 2019.
22. B. J. Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):2, 2002.
23. G. Gigerenzer. *How to stay smart in a smart world: Why human intelligence still beats algorithms*. Penguin UK, 2022.
24. H. W. Gottinger and P. Weimann. Intelligent decision support systems. *Decision support systems*, 8(4):317–332, 1992.
25. B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.
26. A. Grigoryan. "you are too blunt, too ambitious, too confident": Cultural messages that undermine women's paths to advancement and leadership in academia and beyond. In *Surviving Sexism in Academia*, pages 243–249. Routledge, 2017.
27. N. Hanna, D. Richards, et al. Do birds of a feather work better together? the impact of virtual agent personality on a shared mental model with humans during collaboration. In *COOS@ AAMAS*, pages 28–37, 2015.
28. H. Hester. Technology becomes her. *New Vistas*, 3(1):46–50, 2017.
29. M. Jain, P. Kumar, R. Kota, and S. N. Patel. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference*, pages 895–906, 2018.
30. M. T. Johnson and A. H. Vera. No ai is an island: The case for teaming intelligence. *AI Mag.*, 40:16–28, 2019.

31. E. H. Jung, T. F. Waddell, and S. S. Sundar. Feminizing robots: User responses to gender cues on robot body and screen. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3107–3113, 2016.
32. M. Kang. A study of chatbot personality based on the purposes of chatbot. *The Journal of the Korea Contents Association*, 18(5):319–329, 2018.
33. H. Kim, D. Y. Koh, G. Lee, J.-M. Park, and Y.-k. Lim. Designing personalities of conversational agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2019.
34. Y. Kim, A. L. Baylor, and E. Shen. Pedagogical agents as learning companions: the impact of agent emotion and gender. *Journal of Computer Assisted Learning*, 23(3):220–234, 2007.
35. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.
36. D. E. Lee. Ideal female-male traits and evaluation of favorability. *Perceptual and Motor Skills*, 50(3_suppl):1039–1046, 1980.
37. N. Lessio and A. Morris. Toward design archetypes for conversational agent personality. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3221–3228. IEEE, 2020.
38. T. W. Liew and S.-M. Tan. Social cues and implications for designing expert and competent artificial agents: A systematic review. *Telematics and Informatics*, 65:101721, 2021.
39. T. W. Malone. How can human-computer "superminds" develop business strategies? *The Future of Management in an AI World*, 2019.
40. M. McDonnell and D. Baxter. Chatbots and gender stereotyping. *Interacting with Computers*, 31(2):116–121, 2019.
41. B. Mehra. Chatbot personality preferences in global south urban english speakers. *Social Sciences & Humanities Open*, 3(1):100131, 2021.
42. P. Nag and Ö. N. Yalçın. Gender stereotypes in virtual agents. In *Proceedings of the 20th ACM International conference on intelligent virtual agents*, pages 1–8, 2020.
43. C. Nass, Y. Moon, and N. Green. Are machines gender neutral? gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, 27(10):864–876, 1997.
44. N. J. Nilsson. *The quest for artificial intelligence*. Cambridge University Press, 2009.
45. H. Noble and G. Mitchell. What is grounded theory? *Evidence-based nursing*, 19(2):34–35, 2016.
46. D. A. Norman. *Emotional design: Why we love (or hate) everyday things*. Civitas Books, 2004.
47. A. Parker and J. Tritter. Focus group method and methodology: current practice and recent debate. *International Journal of Research & Method in Education*, 29(1):23–37, 2006.
48. J. W. Pennebaker and L. A. King. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296, 1999.
49. G. Phillips-Wren, M. Mora, G. A. Forgionne, and J. N. Gupta. An integrative evaluation framework for intelligent decision support systems. *European Journal of Operational Research*, 195(3):642–652, 2009.
50. J.-C. Pomerol. Artificial intelligence and human decision making. *European Journal of Operational Research*, 99(1):3–25, 1997.

51. M. Porcheron, J. E. Fischer, S. Reeves, and S. Sharples. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.
52. A. Rapp, L. Curti, and A. Boldi. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies*, 151:102630, 2021.
53. Q. Roy, M. Ghafurian, W. Li, and J. Hoey. Users, tasks, and conversational agents: A personality study. In *Proceedings of the 9th International Conference on Human-Agent Interaction*, pages 174–182, 2021.
54. E. Ruane, A. Birhane, and A. Ventresque. Conversational ai: Social and ethical considerations. In *AICS*, pages 104–115, 2019.
55. E. Ruane, S. Farrell, and A. Ventresque. User perception of text-based chatbot personality. In *International Workshop on Chatbot Research and Design*, pages 32–47. Springer, 2020.
56. E. Ruane, S. Farrell, and A. Ventresque. User perception of text-based chatbot personality. In *International Workshop on Chatbot Research and Design*, pages 32–47. Springer, 2021.
57. S. J. Russell and P. Norvig. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.
58. L. Sanny, A. Susastra, C. Roberts, and R. Yusramdaleni. The analysis of customer satisfaction factors which influence chatbot acceptance in indonesia. *Management Science Letters*, 10(6):1225–1232, 2020.
59. C. Shani, A. Libov, S. Tolmach, L. Lewin-Eytan, Y. Maarek, and D. Shahaf. "alexa, do you want to build a snowman?" characterizing playful requests to conversational agents. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.
60. R. Sharda, S. H. Barr, and J. C. McDonnell. Decision support system effectiveness: a review and an empirical test. *Management science*, 34(2):139–159, 1988.
61. B. A. Shawar and E. S. Atwell. Using corpora in machine-learning chatbot systems. *International journal of corpus linguistics*, 10(4):489–516, 2005.
62. T. L. Smestad and F. Volden. Chatbot personalities matters. In *International Conference on Internet Science*, pages 170–181. Springer, 2018.
63. L. R. Soenksen, T. Kassis, S. T. Conover, B. Marti-Fuster, J. S. Birkenfeld, J. Tucker-Schwartz, A. Naseem, R. R. Stavert, C. C. Kim, M. M. Senna, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581):eabb3652, 2021.
64. L. G. Terveen. Overview of human-computer collaboration. *Knowledge-Based Systems*, 8(2-3):67–81, 1995.
65. L. Vanderlyn, G. Weber, M. Neumann, D. Väth, S. Meyer, and N. T. Vu. "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57, 2021.
66. T. Verhagen, J. Van Nes, F. Feldberg, and W. Van Dolen. Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, 19(3):529–545, 2014.
67. S. T. Völkel and L. Kaya. Examining user preference for agreeableness in chatbots. In *CUI 2021-3rd Conference on Conversational User Interfaces*, pages 1–6, 2021.
68. S. T. Völkel, R. Schoedel, L. Kaya, and S. Mayer. User perceptions of extraversion in chatbots after repeated use. In *CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

69. L. Wang, D. Wang, F. Tian, Z. Peng, X. Fan, Z. Zhang, M. Yu, X. Ma, and H. Wang. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–31, 2021.
70. H. Xiao, D. Reid, A. Marriott, and E. Gulland. An adaptive personality model for ecas. In *International Conference on Affective Computing and Intelligent Interaction*, pages 637–645. Springer, 2005.
71. J. Xiao, J. Stasko, and R. Catrambone. The role of choice and customization on users' interaction with embodied conversational agents: effects on perception and performance. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1293–1302, 2007.
72. Z. Xiao, M. X. Zhou, and W.-T. Fu. Who should be my teammates: using a conversational agent to understand individuals and help teaming. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019.
73. M. X. Zhou, C. Wang, G. Mark, H. Yang, and K. Xu. Building real-world chatbot interviewers: Lessons from a wizard-of-oz field study. In *IUI Workshops*, 2019.