10-9-2023

A Picture is Worth a Thousand Words – Exploring Bias in Inclusive Chatbot Design

Ricarda Schlimbach
Technische Universität Braunschweig, Germany, r.schlimbach@tu-braunschweig.de

Anika Stoppel
Technische Universität Braunschweig, Germany, a.stoppel@tu-bs.de

Lucas Lampka
Technische Universität Braunschweig, Germany, l.lampka@tu-bs.de

Susanne Robra-Bissantz
Technische Universität Braunschweig, Germany, s.robra-bissantz@tu-bs.de

A Picture is Worth a Thousand Words – Exploring Bias in Inclusive Chatbots

Research Paper

Ricarda Schlimbach[1,2], Anika Stoppel[1], Lucas Lampka[1], and Susanne Robra-Bissantz[1]

[1] TU Braunschweig, Chair of Service Information Systems, Braunschweig, Germany
{r.schlimbach, a.stoppel, l.lampka, s.robra-bissantz}@tu-bs.de
[2] Hochschule Heilbronn, Wirtschaftsinformatik, Schwäbisch Hall, Germany

Abstract. This study examines the impact of different avatar pictures (gender & disability representation) and gendering on students' perceptions of chatbots in an interaction on learning strategies with 180 students from a German university. In the first experiment, we manipulated the chatbot's humanoid profile picture based on gender and the representation of a visible handicap (wheelchair). In the second experiment, we varied its language style. Statistical analysis revealed that displaying a physical disability significantly enhanced trust, credibility, and empathy but reduced perceived competence and dominance. Gender-sensitive language improved perceptions of competence, trust, credibility, and empathy, whereas we did not find significant interaction effects between both factors. Our results imply the necessity of a more inclusive design of information systems and highlight designers' responsibility in raising awareness and mitigating unconscious bias, as digital learning (technologies) continue to advance.

Keywords: Inclusiveness, PCA, bias, chatbot, diversity, social cues.

1 Introduction

The inclusion of individuals with physical or mental disabilities in learning environments at schools and universities should be viewed as a given, not a privilege (Miyauchi, 2020). Inclusion in the context of learning means that all learners, including those with physical or mental disabilities, have the same opportunities to participate in the learning experience (Francisco et al., 2020). However, individuals with physical or mental disabilities are often attributed with less competence; for instance in job application processes, resulting in being less frequently considered for job postings (Bonaccio et al., 2020). Similarly, even young students tend to believe the statements of physically healthy, non-disabled individuals were more credible than those of someone with a physical disability, although they cannot always explain this biased perception that reflects the deeply ingrained prejudices against minorities (Jaffer & Ma, 2015; Manzanero et al., 2015). In addition to the inclusion of people with physical or mental disabilities, mitigating gender bias is also an important issue (Bernagozzi et al., 2021;

Fossa & Sucameli, 2022). Moreover, gender-sensitive language might reduce prejudice and discrimination by integrating everyone equally into language use, especially in learning environments (Formanowicz et al., 2015; Sczesny et al., 2016).

As teaching and learning are increasingly shifting into the digital space and being technologically supported, socially inclusive design of information systems for educational purposes must be taken into greater consideration (Schlimbach & Robra-Bissantz, 2022), especially because humans tend to apply social rules, norms, expectations, and attitudes from reality to situations in which they communicate with machines (Lee & Nass, 2010). In this context, the inclusive design of Pedagogical Conversational Agents (PCAs) appears to be particularly relevant, as they embody social cues such as a human identity (e.g., avatar, gender) and verbal cues (Feine et al., 2019; Seeger et al., 2018) like natural and potentially gender-sensitive language, while also gaining attention in research and practice (Khosrawi-Rad et al., 2022). Despite several studies that target minorities as the key user group for interacting with PCAs such as disadvantaged learners (Gupta & Chen, 2022) or internationals with language barriers (Schlimbach, Khosrawi-Rad, et al., 2022), research that explores the impact of their inclusive design is scarce to find. PCAs neither represent visually disabled people nor do they use gender-sensitive language to foster diversity in education (Schlimbach & Robra-Bissantz, 2022). This seems surprising, in that diversity representation has been found to positively influence students' respect towards minorities and results essential for inclusiveness in education (Al-Obaydi, 2019; Holoien, 2013; Lie et al., 2021). Understanding inclusive PCA design and how it leads to (potentially biased) responses of the PCA's characteristics is thus of high practical relevance. We, therefore, intend to contribute to the following research questions (RQ):

To what extent does the visual portrayal of a physical disability of a PCA (avatar portrayed with a visible physical disability) influence students' perception of (a) its competence, (b) trust, (c) credibility, (d) dominance, and (e) empathy ***(RQ1)****? To what extent do the results found in (a) to (e) differ between a female and a male PCA's representation* ***(RQ2)****? How does the use of gender-sensitive language by a PCA influence users' perception of the variables (a) to (e) mentioned before* ***(RQ3)****?*

Findings on how the portrayal of physical disability and the use of gender-sensitive language in PCAs affect students' perception might build upon recent findings (Fossa & Sucameli, 2022; Schlimbach & Robra-Bissantz, 2022) and create further awareness for the mindful and socially inclusive design of supportive tools (in education).

2 Research Background

2.1 Designing Inclusive PCAs

PCAs facilitate learning by offering interactive support to students in their role of a tutor, peer student, or coach (Hobert & Meyer von Wolff, 2019) and mitigate learning content (Weber et al., 2021). In interacting, human users attribute social qualities to their artificial conversation partners according to the computers-are-social-actors theory (CASA), even when they know that they are interacting with a computer (Lee & Nass, 2010). Adapted social cues (Seeger et al., 2018) can intentionally design PCAs

to promote inclusiveness in educational technology (Lembcke et al., 2020), and thus, influence human perception in real life (Feine et al., 2020). For instance, a PCA's human identity can be embodied by its name (Cowell & Stanney, 2005), gender (Schuetzler et al., 2018), and avatar representation (Gong, 2008), while emojis transmit nonverbal cues (Hu et al., 2018) and visual design (Gupta & Chen, 2022). Verbal cues, such as language style (Gnewuch et al., 2018), self-references (Schuetzler et al., 2018), and personal introduction (Cafaro et al., 2016), help circumscribe the category of cues. Each of these elements may potentially promote inclusiveness by capturing underrepresented characteristics (Holoien, 2013; Ydo, 2020). Inclusion begins with the individual perception of differences between people (Boban & Hinz, 2009). Therefore, it is of enormous importance to sensitize understanding and to establish inclusive structures both in social life and information systems.

2.2 Gender-neutral Language

The discussion about the introduction of gender-neutral language has been ongoing in Germany since the 1970s. It deals with the fact that in German grammar, the generic masculine is used when describing a group of both female and male individuals. This means that a group of female and male students is referred to as "Studenten," which implies that both female and male individuals are meant equally. However, this becomes problematic when using the generic masculine causes cognitive exclusion of female individuals (“Studentinnen”) and results in their being given less attention in various contexts (Braun et al., 2007). This phenomenon has been confirmed in prior studies (e.g., Heise, 2000) and resulted in alternatives to have all genders represented in language use, including neutralization, using both gender forms, or adding symbols like an asterix (*) to represent inclusiveness (Braun et al., 2007).
With the introduction of § 1 Abs. 2 of the Federal Equality Act, which obliges the use of gender-equitable language, the discussion about the necessity of moving away from the generic masculine has shifted to how and in what form the implementation should take place. However, there is also a controversial debate on the use of gender-sensitive language, as it decreases comprehensibility, harms grammatical correctness and thus might set further language barriers (Braun et al., 2007; Chubb & Derrick, 2020).

2.3 Theoretical Foundations that Lead to Bias

Social identity theory (Tajfel & Turner, 1979), focuses on how individuals derive their sense of self and social identity from group membership. It suggests that people tend to categorize themselves and others into social groups and develop a sense of belonging and self-esteem based on their group membership. The theory also addresses the concepts of in-group favoritism, where individuals show a preference for their own group, and out-group aversion, where individuals may exhibit negative attitudes or behaviors towards members of other groups. In addition, the Stigma theory, developed by Erving Goffman in 1963, examines the social process of labeling and devaluing individuals or groups based on perceived differences or deviations from societal norms. It explores how stigmatized individuals may experience social exclusion, discrimination, and the

negative effects of stereotypes. A stigma is a usually visible attribute that sets individuals apart from others. For example, this could be skin color or a physical disability. The result of a stigma could be that the affected individual is not being fully accepted within social situations (Goffman, 1986).
When people first meet, they tend to perceive each other based on their most prominent characteristics. This phenomenon is known as the *Halo Effect*, when positive characteristics or features lead to a positive assessment of attributes (Landy & Sigall, 1974), or as the *Horn Effect*, when negative characteristics or features lead to a negative assessment of attributes (Pohl, 2016; Shifrer, 2013). Often, other attributes are being overlooked (Domsch et al., 2019). For instance, in the context of this research project, people with physical or mental disabilities are often judged primarily based on this characteristic and are then perceived as less competent in general, even in situations where the handicap is irrelevant (Fiske et al., 2007; Rohmer & Louvet, 2018). People have preconceived notions of social groups and their behaviors and abilities, which are unreflectively applied to every member of that group. Stereotypes, unlike prejudices, are not emotional but rather neutral. Prejudices are internalized and individually confirmed stereotypes that people believe to be true. Bias like gender bias (Oh et al., 2019) or affinity bias (Trainer et al., 2020) in this context refers to perception distortions that arise, among others, from prejudices, sometimes even arising unconsciously. These biases are evolutionarily determined to quickly process information to survive. For instance, affinity bias makes people prefer others who are similar to themselves or to close associates, while beauty bias associates a person's appearance with certain qualities (Domsch et al., 2019). Gender bias provokes distortions due to the perceived gender of a person (Striebing, 2021), whereas confirmation bias describes people searching for confirming information that supports their thought patterns (Domsch et al., 2019). Due to the CASA theory (Nass & Moon, 2000), these biases might translate to the perception of PCAs and initial findings prove that gender bias, affinity bias, and unconscious bias towards disabilities are shown towards chatbots as well (Feine et al., 2020; Schlimbach & Robra-Bissantz, 2022; Zabel & Otto, 2021).

3 Research Design

To understand the relationship between the representation of physical disability and the use of gender-neutral language and the target variables of competence, trust, credibility, dominance, and empathy as variables relevant to learning success (cf. following sections), we designed six PCAs and derived 15 hypotheses to be tested in experiments.

3.1 Deriving Hypotheses

Perceived Competence. The perception of a person's competence in a given activity is determined by the successful integration of their accumulated knowledge, skills, and experience relevant to the situation (Fiske et al., 2007; Schmid et al., 2022). In our experiment, we evaluate the perceived domain expertise of the PCAs in teaching learning strategies. As all PCAs possess the same subject matter knowledge and response

repertoire, their expertise is identical. According to studies, 77 percent of chatbot users expect the technical competence to be at least equivalent to, or even more advanced than that of a human (Furchheim et al., 2021). Individuals with physical or mental disabilities are perceived as less competent in real life (Fiske et al., 2007). Women also appear less competent than men (Oh et al., 2019), whereas users of gender-sensitive language appear more competent in a study than those who do not use it (Hannover & Wolter, 2019; Vervecken & Hannover, 2012). We anticipate that these findings transfer to PCAs. **H1**: The portrayal of a physical handicap has a negative impact on perceived competence. **H2**: Using a female profile picture has a negative impact on the PCA's perceived competence. **H3**: The use of gender-sensitive language has a positive impact on perceived competence.

Trust. Schweer (1996) and Luhmann (1989) describe trust as a fundamental attitude that a person holds towards others. The success of an educational intervention (Schulte-Pelkum et al., 2014) is positively correlated with the level of trust established between the tutor and student, and consequently between the PCA and its user (Schweer, 2000). Women are perceived as more trustworthy than men in general (Buchan et al., 2008). Regarding trust in people with physical or mental disabilities, no evidence was found in the literature so we assume this trait has no impact on trust. People also tend to trust initiatives that use gender-sensitive language more, while those that do not are viewed as less competent and less credible (Formanowicz et al., 2015). The same applies to the use of gender-sensitive language. We thus hypothesize **H4**: Depicting physical disability has a positive impact on perceived trust. **H5**: Using a female profile picture has a positive impact on perceived trust. **H6**: Using gender-sensitive language has neither a positive nor negative impact on perceived trust.

Credibility. The variable credibility can be related both to the persons themselves and to the statements made by a person. A person is considered credible when they convey information to another person and believe it to be true, with no intention to deceive the other person. The actual truthfulness of a statement is not a necessary condition for its credibility (Köhnken, 1999). There were no indications found that gender or a physical limitation have an impact on perceived credibility, but Formanowicz et al. (2015) found people not using gender-sensitive language are seen as less credible, which is why we assume the following hypotheses: **H7**: Depicting a physical limitation has no positive or negative influence on perceived credibility. **H8**: The gender displayed in the PCA's profile picture has no positive or negative influence on its perceived credibility. **H9**: The use of gender-sensitive language has a positive influence on perceived credibility.

Dominance. Dominant individuals are described as those who take responsibility, influence and control others, enjoy being in the spotlight, and are willing to assert their goals with aggression if necessary (Cheng et al., 2013). A teacher's dominant behavior has a negative impact on learners' motivation to learn (Jiang & Zhang, 2021). Typically, the trait of dominance is associated more with male individuals (Wirtz, 2021), which is why they are more likely to be perceived as dominant. As described above, inclusion means that society and structures must adapt so that people with and without physical or mental disabilities can operate equally. However, since inclusion has not yet penetrated all areas, people with disabilities still have to adapt to their environment and show

consideration (Felder, 2017). Concerning the perceived dominance in the use of gender-sensitive language, we hypothesize that the gendering chatbot will be perceived as less dominant because it integrates and includes all users and appears more at eye level. We thus hypothesize **H10**: The representation of a physical limitation has a negative impact on perceived dominance. **H11**: The use of a female profile picture for the PCA has a negative impact on perceived dominance. **H12**: The use of gender-sensitive language in a PCA has a negative impact on perceived dominance.

Empathy. Empathy is ascribed to individuals who can empathize with the feelings of others (Light, 2019). In the learning environment, empathy has a positive impact on the learning success of the students as an empathetic teacher motivates them stronger (Tausch, 2008). Women are attributed with a more pronounced ability to empathize than men (Greenberg et al., 2018). As described above, people with physical or mental limitations often have to adapt more frequently in their everyday lives, so it is assumed that they deal with their fellow human beings more empathetically due to their own experiences. The argumentation with regard to gender-inclusive language and its influence on the perception of empathy follows that of the dominance variable. The resulting hypotheses are **H13**: The portrayal of a physical limitation has a positive influence on perceived empathy. **H14**: The use of a female profile picture for the PCA has a positive influence on perceived empathy. **H15**: The use of gender-inclusive language in a PCA has a positive influence on empathy ascribed to the PCA.

3.2 Conducting the Experiment

Within the framework of this research project, the RQs were investigated in two experiments designed between subjects with a total of 180 participating students. We anticipated a medium effect size (Cohen's $d = 0.6$), a desired statistical power level of 0.8 and a probability level of $p < 0.05$. The a priori power analysis (Soper, 2023) thus resulted in a minimum total sample size (two-tailed hypothesis) of 90 per experiment. This approach is complemented by a quantitative survey with qualitative open questions to leverage the strengths of each method and achieve informative results (Halcomb & Hickman, 2015).

First, we created the PCAs 1-4 with *Google Dialogflow* with a consistent communication style but manipulated their profile pictures as shown in **Figure *1***. PCA 1 is represented by Ben, a male person without an apparent disability. PCA2 embodies again Ben but displays his physical limitation (a wheelchair). PCA3 shows Lena, a female without apparent disability, while PCA4 represents her with physical limitations (visible wheelchair). Second, we manipulated the gender-sensitive language style. Max uses gender-sensitive language embedded into PCA5, while PCA6 interacts using the generic masculine. The conversation content is identical for all six PCAs, as we provided pre-defined response buttons for users to select from, making them less error-prone and ensuring comparability of all conversations. The topic taught by all chatbots is learning strategies, as it is relevant in the learning context cross-disciplinary. Following a between-subject design, we instructed participants to test only one of the six PCAs by randomly assigning them to one PCA depending on their birth month.

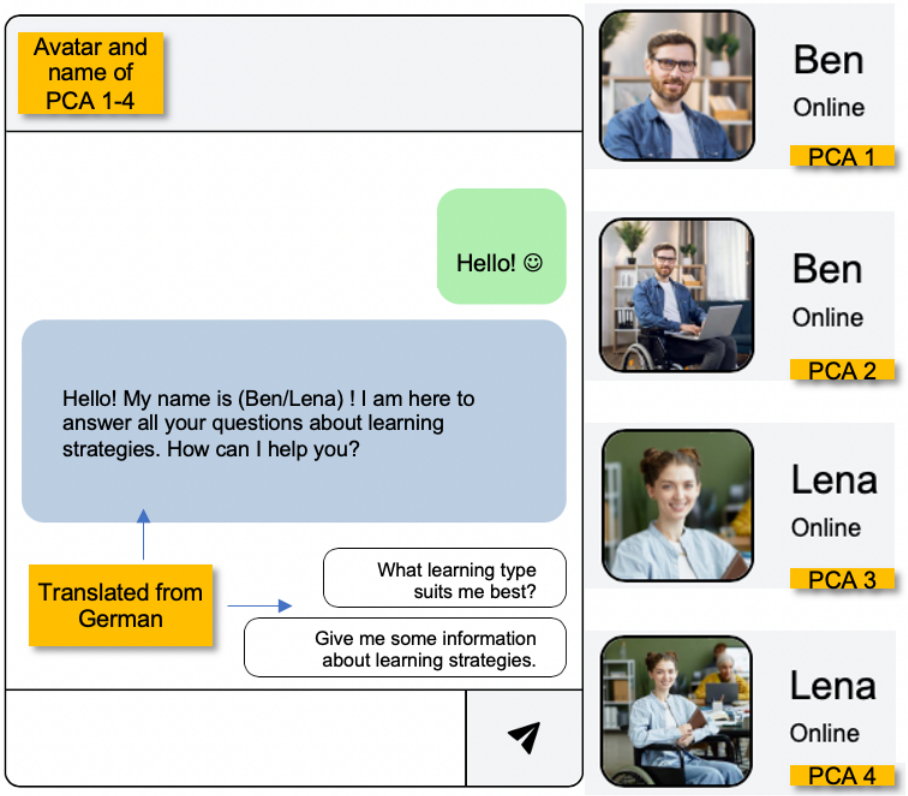

Figure 1. Visualized PCA Manipulations

Thirty participants were recruited per PCA, resulting in a total of 180 participants for the study. The probands were students, as the overall goal was to use a PCA as a learning tutor at university. We initially instructed students on how to interact with the PCA. We monitored that they tested the PCA sufficiently before having them answer a follow-up survey. It mainly queries items on a Likert scale from 1 to 7 ("strongly agree" to "strongly disagree") to assess their perceptions towards the five target constructs to be then statistically analyzed in Jamovi 2.3.21. More precisely, we measured competence (Toader et al., 2020), trust (Simmel et al., 1992; Toader et al., 2020), credibility (Wirth, 1999), empathy (Bodrunova et al., 2019; Früh & Wünsch, 2009), and dominance (Gay, 2000) with scientifically validated items adapted to the use cases of a PCA that teaches learning strategies. Additionally, the survey queried students about attitudes toward gender-sensitive language in chatbots. Finally, we collected sociodemographic data (age, gender, study level).

4 Results

Among the 180 students, 31% are female and 69% male. 47.8% of all participants had already tried a chatbot in a learning context, while 52.2% of the participants did not. The majority of the participants was pursuing a bachelor's degree, 35% were studying towards a master's degree, and 22 had completed their master's degree, among these 1,1% completed a PhD, or graduated from vocational training (1,1%) before. There was a great diversity of majors covered in this study such as Information Systems, Technology Management, Psychology or Engineering.

Demographic data per PCA group (e.g., group 1 tested PCA1) concerning gender, age and study level (Bachelor's (BSc); Master's (MSc) or PhD Students (PhD)) are summarized in Table 1.

Table 1. Demographic Data

	Gender		Age Group				Study Level		
	male	**female**	**< 20**	**20-24**	**25-29**	**> 29**	**BSc**	**MSc**	**PhD**
Group 1	17	13	5	12	8	5	11	13	6
Group 2	21	9	9	11	8	2	12	12	6
Group 3	18	12	5	17	5	3	15	12	3
Group 4	17	13	6	12	9	3	18	8	4
Group 5	25	5	10	14	3	3	23	6	1
Group 6	26	4	10	13	6	1	16	12	2
	124	56	45	79	39	17	95	63	22

4.1 Quantitative Results

First, we performed a reliability analysis in Jamovi to check for internal construct validity. Then, we conducted an exploratory factor analysis to verify that the items correspond to their respective constructs. For each of the five constructs (competence, trust, credibility, empathy, and dominance), a Cronbach's alpha was calculated to assess its reliability. These alpha values were found to be good or even excellent ranging from 0.842 (credibility) to 0.955 (competence). The exploratory factor analysis confirmed that the items were correctly assigned to their respective constructs. Since the Shapiro-Wilk normality test (p (competence, trust, empathy, dominance) $< .001$; p (credibility) $= 0.009$) and Levene's test of homogeneity ($p = 0.157$)) were violated, the use of Student's t-test rendered inappropriate. Therefore, we picked the Welch's t-test for independent samples fur further analysis.

Impact on visible physical disability and gender. The comparison between the portrayal of physical disability (PCA 2&4) and no visible disability (PCA 1&3) is summarized in Table 2 and depicts the results of the Welch's t-test (non-disabled vs. disabled) with 120 participants that judged on the chatbots (between subject design).

Table 2. Welch's t-Test Comparing Disabled vs. non-Disabled Avatar Pictures

Non-disabled vs. disabled		t	df	p		Effect size
Competence	Welch's t	2.13	103	0.036	Cohens d	0.388
Trust	Welch's t	-1.78	114	0.078	Cohens d	-0.325
Credibility	Welch's t	- 3.34	110	0.001	Cohens d	-0.609
Empathy	Welch's t	-15.72	118	< .001	Cohens d	-2.870
Dominance	Welch's t	5.13	112	< .001	Cohens d	0.937

In summary, the results show that **not**[1] depicting Ben's and Lena's physical disability had a small positive effect on the competence variable (Cohen's D (CD) = 0.388), a small negative effect on trust (CD = - 0.325), a medium negative effect (CD = - 0.609) on credibility, and a large negative effect on empathy (CD = - 2.87). On the other hand, the perception of dominance was affected largely positively (CD = 0.937). Table 3 additionally presents the descriptive statistics of the variables under investigation.

Table 3. Descriptive Statistics on Disabled vs. Non-Disabled PCAs in Comparison

	PCA	Comp	Trust	Cred	Empa	Domi
Mean	1&3	5.91	5.43	5.27	3.47	4.36
	2&4	5.50	5.66	5.68	5.91	3.49
Standard Deviation	1&3	1.22	0.65	0.58	0.85	0.81
	2&4	0.82	0.78	0.77	0.85	1.03

For the analysis and answering of research questions, it is also of interest whether the gender of the avatar has an influence on the dependent variables. We thus conducted another Welch's t-Test comparing both PCAs' representing Ben with the other two showing Lena. The comparison shows again significant differences:

Table 4. Welch's t-Test Comparing Avatar Pictures of Ben vs. Lena

Ben 1+2 vs. Lena 3+4		t	df	p		Effect size
Competence	Welch's t	-2.200	117	0.030	Cohens d	-0.402
Trust	Welch's t	-6.019	117	< .001	Cohens d	-1.099
Credibility	Welch's t	-3.048	115	0.003	Cohens d	-0.556
Empathy	Welch's t	-0.600	116	0.550	Cohens d	-0.110
Dominance	Welch's t	-2.384	118	0.019	Cohens d	-0.435

As depicted, the female representation has a small negative effect (CD = - 0.402) on perceived competence. So, Lena is perceived significantly less competent than Ben, although both PCAs mediate the same factual knowledge and differ exclusively regarding their profile picture. Besides, the female representation has a large negative effect on trust (CD = - 1.099), a medium negative effect on credibility (CD = - 0.556), and a

[1] The initial publication of this paper in the AIS eLibrary failed to include the negation in this sentence and thus inverted the statement.

small negative effect on empathy (CD = - 0.11). Moreover, the female representation reveals a medium negative effect on the dominance variable (CD = - 0.435).

By conducting a factorial ANOVA analysis, we found no significant interaction effects for either of the five measured constructs between the two factors of gender and (dis)ability.

Impact on gendering. Next, we compare the results between PCA5 and PCA6, which represent the use of gender-inclusive language and the usage of the generic masculinum for PCA Max, respectively. As Table 5 shows, using gender-inclusive language has a moderate positive effect (CD = 0.601) on the variable of competence. Similarly, the effect is also moderate (CD = 0.673) on the variable of credibility. The use of gender-inclusive language has a large positive effect on the variable of empathy (CD = 1.024). Also, the use of gender-inclusive language indicates a large negative effect (CD = - 1.166) on perceived dominance in our experiment.

Table 5. Welch's t-Test Comparing Gender-sensitive with Generic Language

PCA 5 vs. PCA 6		t	df	p		Effect size
Competence	Welch's t	1.34	47	0.019	Cohens d	0.346
Trust	Welch's t	2.33	50	0.024	Cohens d	0.601
Credibility	Welch's t	2.61	49	0.012	Cohens d	0.673
Empathy	Welch's t	3.97	39	< .001	Cohens d	1.024
Dominance	Welch's t	-4.52	58	< .001	Cohens d	-1.166

We could not control for gender, as PCA 5 and PCA6 showed both the same male person (Max) in the portrayal of the chat interface.

4.2 Qualitative Results

Both PCAs with and without displaying physical limitations appear competent to users because they provide supportive information on learning methods and offer video material beyond the textual conversation, guide the conversation, give detailed answers, and speak error-free. Regarding the comparison of Ben with/without the use of gender-inclusive language, fewer responses were given, but it can be seen that the PCA using gender-inclusive language appears competent because it *"visually exudes confidence"* and provides correct and detailed information. All chatbots were noted as not credible as human tutors, but the credibility of the information provided was confirmed. Students claim that PCAs displaying physical limitations appear empathetic since they embody inclusiveness and *"want others not to feel sorry about their disability"*. In contrast, the PCAs without physical limitations and those without using gender-inclusive language do appear less empathetic as they are perceived more egocentric and the conversation feels *"more mechanic"*. PCAs without displaying physical limitations appear dominant since they control the tone and *"do not allow users to steer the conversation"*.

5 Discussion

In this study, we investigated the effect of different avatar characteristics (gender, physical disability) and the usage of a gender-inclusive language style on six DVs.

Below, in Table 6, we summarize the confirmation (green) or rejection (red) of our hypotheses and document the **eventually measured relations** in our experiment.

Table 6. Table on Experimental Results

H1	The presentation of a physical limitation has a small effect size (Cohen's d = 0.388) and a significant negative impact (p = 0.036) on perceived competence.
H2	The use of a female gender avatar also has a small effect size (Cohen's d = 0.402) and a significant negative impact (p = 0.03) on perceived competence.
H3	The use of gender-inclusive language has a small effect size (Cohen's d = 0.346) and a positive but non-significant (p = 0.186) impact on perceived competence.
H4	The embodiment of a physical limitation has a small effect size (Cohen's d = 0.325) and a significant positive impact (p = 0.078) on perceived trust.
H5	The use of a female avatar image has a large effect size (Cohen's d = 1.099) and a highly significant (p < 0.001) negative impact on perceived trust.
H6	The use of gender-inclusive language has a medium effect size (Cohen's d = 0.601) and a significant (p = 0.024) positive impact on perceived trust.
H7	The embodiment of a physical limitation has a medium effect size (Cohen's d = 0.609) and a highly significant (p = 0.001) positive impact on perceived credibility.
H8	The use of a female avatar also has a medium effect size (Cohen's d = 0.556) and a highly significant (p = 0.003) negative impact on perceived credibility.
H9	The use of gender-inclusive language has a medium effect size (Cohen's d = 0.673) and a significant (p = 0.012) positive impact on perceived credibility.
H10	The presentation of a physical limitation has a large effect size (Cohen's d = 0.939) and a highly significant (p < 0.001) negative impact on perceived dominance.
H11	The use of a female avatar image has a small effect size (Cohen's d = 0.435) and a significant (p = 0.019) negative impact on perceived dominance.
H12	Gender-inclusive language has a large effect size (Cohen's d = 1.166) and a highly significant (p < 0.001) negative impact on perceived dominance.
H13	The portrayal of a physical limitation has a very large effect size (Cohen's d = 2.870) and a highly significant (p < 0.001) positive impact on perceived empathy.
H14	The influence of a female avatar image has no statistically significant effect on perceived empathy (p = 0.550).
H15	The use of gender-inclusive language has a large effect size (Cohen's d = 1.024) and a highly significant (p < 0.001) positive impact on perceived empathy.

The physical world abounds with biases towards people with disabilities, gender, and affinity (Fiske et al., 2007; Oh et al., 2019; Rohmer & Louvet, 2018; Trainer et al., 2020). Our study builds on initial evidence from Schlimbach & Robra-Bissantz (2022) that these biases also apply to virtual information systems like PCAs, consistent with CASA and social response theory (Nass & Moon, 2000). The implications of these findings are relevant for the design of PCAs, chatbots, and other information systems, as inclusive design and its implications need to be considered from the outset.

Although positive discrimination in the virtual world offers the potential to represent diversity characteristics in avatar design, practical implementation is still limited (Schlimbach, Rinn, et al., 2022). However, PCAs can facilitate dialogic encounters between minorities, address discrimination, and contribute to the reflection on diversity (Gupta & Chen, 2022; Lie et al., 2021). Our findings suggest that avatars should be designed carefully to enhance the perceived effectiveness of the PCA. For example,

empathy has a positive effect on learners' success, but dominant behavior by a teacher has a negative impact on learners' motivation (Jiang & Zhang, 2021; Tausch, 2008). However, designing only toward desirable learning outcomes risks excluding diversity and perpetuating unconscious bias. We suggest increasing the visibility of different genders and avatars with physical disabilities in the PCA landscape to enhance diversity awareness and mitigate bias.

Our research implications are significant for designing and developing PCAs and other virtual information systems. By highlighting the impact of avatar design on users' perceptions, our findings underscore the importance of inclusive design practices from the outset. Designers and developers should consider the representation of diverse genders and visible physical disabilities in the PCA landscape to foster diversity awareness and mitigate bias. Furthermore, our study highlights the potential of PCAs to facilitate dialogic encounters, address discrimination, and promote reflection on diversity. Educators and policymakers can leverage this potential to create inclusive learning environments that promote empathy, trust, and credibility. By integrating inclusive design principles and fostering diversity awareness, PCAs and other information systems can play a crucial role in promoting equity and inclusivity in various contexts.

It is important to acknowledge some limitations of our study. First, our research focused on the visual portrayal of a physical disability and gender-sensitive language in PCAs within an educational context. However, avatars can possess various other characteristics, such as age, race, or clothing, which could also influence users' perceptions. Future research should explore the effects of these additional factors and their interactions to provide a more comprehensive understanding of avatar design and its impact on user perception. Second, our study was conducted with German university students, and the results may not generalize to other populations or cultural contexts. Replicating the experiments with diverse participant groups and in different cultural settings would enhance the external validity of the findings. Moreover, while our research focused on PCAs in educational settings, the implications of avatar design and inclusive language extend beyond education. Further investigations in different domains and target groups will contribute to a more comprehensive understanding of the potential limitations and benefits of diverse avatar representation. In addition, gender-inclusive language showed positive effects on the perceived trust and credibility of our PCAs, but the sample size for that specific t-test was below the calculated 90 probands in the a priori power analysis and should thus be repeated with a larger sample.

In summary, future studies could investigate how different avatar characteristics (e.g., age, race, clothing) impact the perceived effectiveness of a PCA and whether certain groups respond differently to manipulated factors. Mitigating social and algorithmic bias is particularly challenging since underrepresented characteristics and perceptions of diversity are context-specific and diverge widely between cultures, application fields, and target groups (Koopmans & Schaeffer, 2017). Further studies are needed to address this challenge and promote inclusion and diversity in PCAs and other information systems. Designers and developers should mindfully consider social cues (Feine et al., 2019), i.e. gendering and inclusiveness to mitigate bias without harming variables relevant to learning with PCAs.

References

Al-Obaydi, L. (2019). Cultural Diversity, Awareness and Teaching: A Study in an EFL Context. Journal of Asia TEFL, 16, 987–995. https://doi.org/10.18823/asiatefl.2019.16.3.15.987

Bernagozzi, M., Srivastava, B., Rossi, F., & Usmani, S. (2021). Gender Bias in Online Language Translators: Visualization, Human Perception, and Bias/Accuracy Tradeoffs. IEEE Internet Computing, 25(5), 53–63. https://doi.org/10.1109/MIC.2021.3097604

Boban, I., & Hinz, A. (2009). Der Index für Inklusion: Lernen und Teilhabe in der Schule der Vielfalt entwickeln. Sozial Extra, 33(9–10), 12–16. https://doi.org/10.1007/s12054-009-0078-4

Bodrunova, S. S., Koltsova, O., Følstad, A., Halpin, H., Kolozaridi, P., Yuldashev, L., Smoliarova, A., & Niedermayer, H. (Eds.). (2019). Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers (Vol. 11551). Springer International Publishing. https://doi.org/10.1007/978-3-030-17705-8

Bonaccio, S., Connelly, C. E., Gellatly, I. R., Jetha, A., & Martin Ginis, K. A. (2020). The Participation of People with Disabilities in the Workplace Across the Employment Cycle: Employer Concerns and Research Evidence. Journal of Business and Psychology, 35(2), 135–158. https://doi.org/10.1007/s10869-018-9602-5

Braun, F., Oelkers, S., Rogalski, K., Bosak, J., & Sczesny, S. (2007). "Aus Gründen der Verständlichkeit ...": Der Einfluss generisch maskuliner und alternativer Personenbezeichnungen auf die kognitive Verarbeitung von Texten. Psychologische Rundschau, 58(3), 183–189. https://doi.org/10.1026/0033-3042.58.3.183

Buchan, N. R., Croson, R. T. A., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the Investment Game. Journal of Economic Behavior & Organization, 68(3–4), 466–476. https://doi.org/10.1016/j.jebo.2007.10.006

Cafaro, A., Vilhjálmsson, H. H., & Bickmore, T. (2016). First impressions in human-agent virtual encounters. ACM Transactions on Computer-Human Interaction, 23(4), 1–40. https://doi.org/10.1145/2940325

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. Journal of Personality and Social Psychology, 104(1), 103–125. https://doi.org/10.1037/a0030398

Chubb, J., & Derrick, G. E. (2020). The impact a-gender: Gendered orientations towards research Impact and its evaluation. Palgrave Communications, 6(1), Article 1. https://doi.org/10.1057/s41599-020-0438-z

Cowell, A. J., & Stanney, K. M. (2005). Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. International Journal of Human-Computer Studies, 62(2), 281–306. https://doi.org/10.1016/j.ijhcs.2004.11.008

Domsch, M. E., Ladwig, D. H., & Weber, F. C. (2019). UNCONSCIOUS BIAS – Eine Einführung zu Vorurteilen im Arbeitsleben. In M. E. Domsch, D. H. Ladwig, & F. C. Weber (Eds.), Vorurteile im Arbeitsleben (pp. 3–20). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-59232-8_1

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. International Journal of Human-Computer Studies, 132, 138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender Bias in Chatbot Design: Vol. 11970 LNCS (p. 93). Scopus. https://doi.org/10.1007/978-3-030-39540-7_6

Felder, F. (2017). Inklusion und Arbeit: Was steht auf dem Spiel? In C. Misselhorn & H. Behrendt (Eds.), Arbeit, Gerechtigkeit und Inklusion (pp. 99–119). J.B. Metzler. https://doi.org/10.1007/978-3-476-04374-0_7

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. Trends in Cognitive Sciences, 11(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Formanowicz, M. M., Cisłak, A., Horvath, L. K., & Sczesny, S. (2015). Capturing socially motivated linguistic change: How the use of gender-fair language affects support for social initiatives in Austria and Poland. Frontiers in Psychology, 6. https://doi.org/10.3389/fpsyg.2015.01617

Fossa, F., & Sucameli, I. (2022). Gender Bias and Conversational Agents: An ethical perspective on Social Robotics. Science and Engineering Ethics, 28(3). Scopus. https://doi.org/10.1007/s11948-022-00376-3

Francisco, M. P. B., Hartman, M., & Wang, Y. (2020). Inclusion and Special Education. Education Sciences, 10(9), 238. https://doi.org/10.3390/educsci10090238

Früh, W., & Wünsch, C. (2009). Empathie und Medienempathie. Publizistik, 54(2), 191–215. https://doi.org/10.1007/s11616-009-0038-9

Furchheim, P., Monard, F., Uebersax, H.-P., Müller, M., Hannich, F., & Kaya, T. (2021). Chatbot Studie ’21. Institut für Marketing Management. https://www.intre.cc/wp-content/uploads/2021/09/Chatbot-Studie-2021_aiaibot_ZHAW.pdf

Gay, F. (Ed.). (2000). DISG-Persönlichkeitsprofil: Verstehen Sie sich selbst besser , schöpfen Sie Ihre Möglichkeiten aus, entdecken Sie Ihre Stärken und Schwächen; [mit dem Original DISG-Testmaterial zur Selbstauswertung] (16. Aufl., 106.-110 Tsd). DISG-Training.

Gnewuch, U., Morana, S., Adam, M., & Maedche, A. (2018). “The Chatbot is typing …” – The Role of Typing Indicators in Human-Chatbot Interaction.

Goffman, E. (1986). Stigma: Notes on the management of spoiled identity (1. Touchstone ed). Simon & Schuster.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. Computers in Human Behavior, 24(4), 1494–1509. https://doi.org/10.1016/j.chb.2007.05.007

Greenberg, D., Warrier, V., Allison, C., & Baron-Cohen, S. (2018). Testing the Empathizing-Systemizing theory of sex differences and the Extreme Male Brain theory of autism in half a million people. Proceedings of the National Academy of Sciences, 115, 1811032115. https://doi.org/10.1073/pnas.1811032115

Gupta, S., & Chen, Y. (2022). Supporting Inclusive Learning Using Chatbots? A Chatbot-Led Interview Study. Journal of Information Systems Education, 33(1), 98–108.

Halcomb, E., & Hickman, L. (2015). Mixed methods research. Nursing Standard, 29(32), 41–47. https://doi.org/10.7748/ns.29.32.41.e8858

Hannover, B., & Wolter, I. (2019). Geschlechtsstereotype: Wie sie entstehen und sich auswirken. In B. Kortendiek, B. Riegraf, & K. Sabisch (Eds.), Handbuch Interdisziplinäre Geschlechterforschung (pp. 201–210). Springer Fachmedien. https://doi.org/10.1007/978-3-658-12496-0_16

Heise, E. (2000). Sind Frauen mitgemeint? Eine empirische Untersuchung zum Verständnis des generischen Maskulinums und seiner Alternativen. Sprache & Kognition, 19(1/2), 3–13. https://doi.org/10.1024//0253-4533.19.12.3

Hobert, S., & Meyer von Wolff, R. (2019, February 24). Say Hello to Your New Automated Tutor – A Structured Literature Review on Pedagogical Conversational Agents. 14th International Conference on Wirtschaftsinformatik, Siegen, Germany.

Holoien, D. S. (2013). Do Differences Make A Difference? The Effects of Diversity on Learning, Intergroup Outcomes, and Civic Engagement. Princetion University Reports, 18.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., & Akkiraju, R. (2018). Touch Your Heart: A Tone-aware Chatbot for Customer Care on Social Media. 1–12. https://doi.org/10.1145/3173574.3173989

Jaffer, S., & Ma, L. (2015). Preschoolers show less trust in physically disabled or obese informants. Frontiers in Psychology, 5. https://doi.org/10.3389/fpsyg.2014.01524

Jiang, A. L., & Zhang, L. J. (2021). University Teachers' Teaching Style and Their Students' Agentic Engagement in EFL Learning in China: A Self-Determination Theory and Achievement Goal Theory Integrated Perspective. Frontiers in Psychology, 12, 704269. https://doi.org/10.3389/fpsyg.2021.704269

Khosrawi-Rad, B., Rinn, H., Schlimbach, R., Gebbing, P., Yang, X., Lattemann, C., Markgraf, D., & Robra-Bissantz, S. (2022). Conversational Agents in Education – A Systematic Literature Review. ECIS 2022 Research Papers. https://aisel.aisnet.org/ecis2022_rp/18

Köhnken, G. (1999). Glaubwürdigkeit. In R. Lempp, G. Schütze, & G. Köhnken (Eds.), Forensische Psychiatrie und Psychologie des Kindes- und Jugendalters (pp. 318–341). Steinkopff. https://doi.org/10.1007/978-3-662-21913-3_29

Koopmans, R., & Schaeffer, M. (2017). Statistical and Perceived Diversity and Their Impacts on Neighborhood Social Cohesion in Germany, France and the Netherlands. https://doi.org/10.31235/osf.io/64p3z

Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. Journal of Personality and Social Psychology, 29(3), 299–304. https://doi.org/10.1037/h0036018

Lee, J.-E. R., & Nass, C. I. (Eds.). (2010). Trust in Computers: The Computers-Are-Social-Actors (CASA) Paradigm and Trustworthiness Perception in Human-Computer Communication. IGI Global. https://doi.org/10.4018/978-1-61520-901-9.ch001

Lembcke, T.-B., Diederich, S., & Brendel, A. B. (2020). Supporting Design Thinking Through Creative and Inclusive Education Facilitation: The Case of Anthropomorphic Conversational Agents for Persona Building. ECIS 2020 Research Papers. https://aisel.aisnet.org/ecis2020_rp/27

Lie, R., Boogaard, B., & Witteveen, L. (2021). Towards design principles for diversity sensitive learning (pp. 134–146). https://doi.org/10.4324/9781003112525-12

Light, S. N. (2019). The Heterogeneity of Empathy: Possible Treatment for Anhedonia? Frontiers in Psychiatry, 10, 185. https://doi.org/10.3389/fpsyt.2019.00185

Luhmann, N. (1989). Vertrauen: Ein Mechanismus der Reduktion sozialer Komplexität (3., durchgesehene Aufl). F. Enke.

Manzanero, A. L., Alemany, A., Recio, M., Vallet, R., & Aróztegui, J. (2015). Evaluating the Credibility of Statements Given by Persons With Intellectual Disability. [Evaluación de la credibilidad de relatos de personas con discapacidad intelectual]. Anales de Psicología, 31(1), 338–344. https://doi.org/10.6018/analesps.31.1.166571

Miyauchi, H. (2020). A Systematic Review on Inclusive Education of Students with Visual Impairment. Education Sciences, 10(11), 346. https://doi.org/10.3390/educsci10110346

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Oh, D., Buck, E. A., & Todorov, A. (2019). Revealing Hidden Gender Biases in Competence Impressions of Faces. Psychological Science, 30(1), 65–79. https://doi.org/10.1177/0956797618813092

Pohl, R. F. (Ed.). (2016). Cognitive Illusions (0 ed.). Psychology Press. https://doi.org/10.4324/9781315696935

Rohmer, O., & Louvet, E. (2018). Implicit stereotyping against people with disability. Group Processes & Intergroup Relations, 21(1), 127–140. https://doi.org/10.1177/1368430216638536

Schlimbach, R., Khosrawi-Rad, B., & Robra-Bissantz, S. (2022, June 26). Deriving Design Knowledge for eLearning Companions to Support International Students. 35th Bled eConference - Digital Restructuring and Human (Re)Action, Bled, Slovenia. https://doi.org/10.18690/um.fov.4.2022.2

Schlimbach, R., Rinn, H., Markgraf, D., & Robra-Bissantz, S. (2022). A Literature Review on Pedagogical Conversational Agent Adaptation. PACIS 2022 Proceedings. https://aisel.aisnet.org/pacis2022/20

Schlimbach, R., & Robra-Bissantz, S. (2022). Exploring the Impact of Inclusive PCA Design on Perceived Competence, Trust and Diversity. SIGHCI 2022 Proceedings. https://aisel.aisnet.org/sighci2022/9

Schmid, D., Staehelin, D., Bucher, A., Dolata, M., & Schwabe, G. (2022). Does Social Presence Increase Perceived Competence? Evaluating Conversational Agents in Advice Giving through a Vido-Based Survey. 26, ACM Human Computer Interaction(6). https://doi.org/10.1145/3492845

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F. (2018). The Influence of Conversational Agents on Socially Desirable Responding. In Hawaii International Conference on System Sciences 2018 (HICSS-51). https://aisel.aisnet.org/hicss-51/cl/hci/8

Schweer, M. K. W. (1996). Vertrauen in der pädagogischen Beziehung (1. Aufl). Huber.

Schweer, M. K. W. (2000). Vertrauen als basale Komponente der Lehrer-Schüler-Interaktion. In M. K. W. Schweer (Ed.), Lehrer-Schüler-Interaktion (pp. 129–138). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-322-97477-8_6

Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can Gender-Fair Language Reduce Gender Stereotyping and Discrimination? Frontiers in Psychology, 7. https://doi.org/10.3389/fpsyg.2016.00025

Seeger, A.-M., Pfeiffer, J., & Hinzl, A. (2018). Deigning Anthropromorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework.

Shifrer, D. (2013). Stigma of a Label: Educational Expectations for High School Students Labeled with Learning Disabilities. Journal of Health and Social Behavior, 54(4), 462–480. https://doi.org/10.1177/0022146513503346

Simmel, G., Rammstedt, O., & Simmel, G. (1992). Soziologie: Untersuchungen über die Formen der Vergesellschaftung (1. Aufl). Suhrkamp.

Soper, D. (2023). Free A-priori Sample Size Calculator for Student t-Tests—Free Statistics Calculators. https://www.danielsoper.com/statcalc/calculator.aspx?id=47

Striebing, C. (2021, May 18). Karriere-Killer Gender Bias. https://blog.iao.fraunhofer.de/karriere-killer-gender-bias/

Tajfel, H., & Turner, J. C. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worchel (Eds.), The Social Identity Theory of Intergroup Behavior (pp. 33–47). Brooks/Cole Pub. Co.

Tausch, R. (2008). Personzentriertes Verhalten von Lehrern in Unterricht und Erziehung. In M. K. W. Schweer (Ed.), Lehrer-Schüler-Interaktion (pp. 155–176). VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-91104-5_6

Toader, D.-C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., & Rădulescu, A. T. (2020). The Effect of Social Presence and Chatbot Errors on Trust. Sustainability, 12(1), Article 1. https://doi.org/10.3390/su12010256

Trainer, T., Taylor, J., & Stanton, C. (2020, November 18). Choosing the Best Robot for the Job: Affinity Bias in Human-Robot Interaction. https://doi.org/10.1007/978-3-030-62056-1_41

Vervecken, D., & Hannover, B. (2012). Ambassadors of gender equality? How use of pair forms versus masculines as generics impacts perception of the speaker: Perception of speakers of gender fair language. European Journal of Social Psychology, 42(6), 754–762. https://doi.org/10.1002/ejsp.1893

Weber, F., Rüttimann, D., Wambsganss, T., & Söllner, M. (2021, December). Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education.

Wirth, W. (1999). Methodologische und konzeptionelle Aspekte der Glaubwürdigkeitsforschung. In Glaubwürdigkeit in der Marketingkommunikation: Konzeption, Einflussfaktoren und Wirkungspotenzial. (pp. 47–66). Deutscher Universitätsverlag.

Wirtz, M. A. (2021). Maskulinität. In Dorsch Lexikon der Psychologie. https://dorsch.hogrefe.com/stichwort/maskulinitaet

Ydo, Y. (2020). Inclusive education: Global priority, collective responsibility. PROSPECTS, 49(3), 97–101. https://doi.org/10.1007/s11125-020-09520-y

Zabel, S., & Otto, S. (2021). Bias in, Bias Out – the Similarity-Attraction Effect Between Chatbot Designers and Users: Vol. 12764 LNCS (p. 197). Scopus. https://doi.org/10.1007/978-3-030-78468-3_13