A Blueprint for Breaking Stereotypes: Design Principles for Conversational Agents to Empower Girls in Computer Science Education

Completed Research Paper

Daniel Stattkus
German Research Center for Artificial Intelligence
Hamburger Straße 24,
49084 Osnabrück, Germany
daniel.stattkus@dfki.de

Lorena Göritz
German Research Center for Artificial Intelligence
Hamburger Straße 24,
49084 Osnabrück, Germany
lorena.goeritz@dfki.de

Katharina-Maria Illgen
German Research Center for Artificial Intelligence
Hamburger Straße 24,
49084 Osnabrück, Germany
katharina.illgen@dfki.de

Jan Heinrich Beinke
German Research Center for Artificial Intelligence
Hamburger Straße 24,
49084 Osnabrück, Germany
jan.beinke@dfki.de

Oliver Thomas
German Research Center for Artificial Intelligence, Osnabrueck University
Hamburger Straße 24,
49084 Osnabrück, Germany
oliver.thomas@dfki.de

Abstract

In our study, we delve into gender disparities within online computer science courses, focusing on the impact of stereotypes on learners. Our focus is narrowed to gender disparities within online computer science courses, where we investigate the impact of gender stereotypes on learners' choices and performance. Drawing on Stereotype Threat Theory, we pinpoint psychological barriers hindering inclusivity and propose conversational agents as a design intervention to address these challenges. Our conversational agent prototype, developed and evaluated with a seventh-grade class, aims to dismantle gender stereotypes, to motivate girls to pursue computer science, and contribute to broader societal goals of gender equality in the IT field. Utilizing a design science approach, our findings provide actionable insights for platform providers to engage underrepresented users. In addition, our research contributes valuable design knowledge for conversational agents, specifically tailored to support girls in computer science education.

Keywords: Conversational Agent, Stereotype Threat, Online Learning Platform, Gender Equality

Introduction

Over the past few decades, online learning has been an emerging field, with the COVID-19 pandemic accelerating the digital transformation of education as educational institutions were forced to close their doors and offer digital learning options. In some subjects, the distribution of online course participants is biased regarding their demographic characteristics. For example, while careers in science are characterized by a high male quota, more women participate in social domains (National Science Foundation, 2019). One reason for this phenomenon is the prevalence of unconscious expectations based on gender stereotypes (Kerkhoven et al., 2016). Stereotypes are often inaccurate beliefs about the characteristics and behaviors of certain groups that shape societal expectations and influence perceptions about individuals (Devine, 1989). Stereotypical expectations based on fixed norms for women and men, girls, and boys, limit their aspirations, choices, and freedom and, therefore, need to be dismantled (European Commission, 2020). For example, women may find themselves affected by prevailing computer science-based gender stereotypes that assume that computer science is a male profession (Clayton et al., 2009). Because of these stereotypes, girls may find it difficult to relate to STEM fields because they do not align with traditional female gender roles, thus hindering their success in the field (Yücel & Rızvanoğlu, 2019). Steele and Aronson (1995) describe this phenomenon within the Stereotype Threat Theory (STT). Stereotype threat is an individual's anxiety when confronted with negative stereotypes linked to one's identity. This threat can lead to cognitive stress and, consequently, underperformance and fulfillment of the stereotypes (Steele & Aronson, 1995). To address these barriers effectively, it is essential to provide environments, such as learning environments, that reduce the triggering of threats. Designing gender-sensitive online learning environments is a crucial strategy in reducing the impact of stereotype threat (Kizilcec & Saltarelli, 2019). Inclusive design solutions aim to manipulate cues in learning environments to reduce learners' vulnerability to psychological barriers, such as stereotype threat. They aim to provide equal opportunities regardless of learners' sociodemographic background (Kizilcec & Saltarelli, 2019). Inclusive design interventions can contribute to the engagement of an underrepresented group with an online learning platform and manifest in various forms within digital learning contexts (Kizilcec & Saltarelli, 2019). For example, virtual or augmented reality immerses learners in individually stimulating environments, fosters enthusiasm for new subjects, or connects underrepresented groups within a metaverse to strengthen community and identity. However, interventions such as virtual reality pose accessibility and implementation challenges due to hardware requirements, while creating metaverse platforms requires significant resources. Following Faik et al. (2024), who emphasize the critical role of including marginalized communities in technology design to enhance engagement and success, this study explores the use of conversational agents (CAs) to support girls in computer science courses, aiming to directly address and mitigate the unique challenges they face due to prevailing stereotypes. Research suggests that CAs offer an accessible and promising solution for personalized learning experiences that address the barriers marginalized students face on online platforms (Oliveira et al., 2021). This has contributed to the rising relevance of CAs in the global market (Cocchi et al., 2023), as sociotechnical systems — a core focus of Information Systems (IS) research. Additionally, the widespread implementation of CAs is accompanied by an intensified focus on diversity, equity, and inclusion (Abdelhalim et al., 2024). By leveraging existing learning infrastructures, CAs complement and personalize learning content and adapt to different user groups and contexts. Their economic and social impact is expected to rise as individuals, especially vulnerable groups, use CAs for lifelong learning, decision-making, and problem-solving (Karyotaki et al., 2024). Given the potential of CAs as inclusive interventions in digital learning, rigorous research is needed to explore their systematic design for supporting girls in online computer science courses. Research at the intersection of IS and societal challenges of gender inequality in computer science requires a methodologically rigorous approach in line with design science research. Gender-inclusive CAs hold promise for mitigating stereotypes, fostering trust, and reducing unconscious bias by tailoring interactions to support girls and serve as role models (Arroyo et al., 2011). STT from social psychology guides our approach to promoting social inclusion in IS research.

Considering the persistent underrepresentation of women in IT, influenced by various social, cultural, institutional, and individual factors (Botella et al., 2019), our study explores the effectiveness of CAs in male-dominated disciplines to promote inclusivity. We focus on how CAs, when designed with gender-specific cues, can enhance support and act as role models, thus encouraging female participation and identification in IT fields. This research addresses existing contradictions in findings regarding whether opposite gender (Arroyo et al., 2009; Krämer et al., 2016) or matching gender CAs (Arroyo et al., 2011; Rosenberg-Kima et al., 2010) are more effective in supporting users. We synthesized the insights from prior

research in the field of gender-sensitive CA design and refined these insights with expert opinions to formulate design principles (DPs). We thereby respond to the call from Kizilcec and Saltarelli (2019) for more research on how the effectiveness of cues varies based on audience characteristics. By tailoring a CA specifically for girls in computer science through DPs, our study not only contributes to the research avenue "User Characteristics and Adaptive CA Designs" as outlined by (Diederich et al., 2022) but also tackles "Ethical Implications of Designing and Interacting with CAs" by attempting to mitigate gender stereotypes through technological interventions. These considerations lead to the following research question for our study:

How can conversational agents be designed to support girls in online computer science courses?

By applying the design science research approach, we provide the scientific community with design knowledge regarding CAs for gender-sensitive online learning platforms. Our findings enable platform providers to engage users from underrepresented groups with the content of their platform through CA based interventions. By providing gender-sensitive DPs, we address the societal issues of gender inequality by motivating girls to enter the computer science field and breaking stereotypical expectations (European Commission, 2020). The iterative development of the DPs was conducted in four design cycles, following the methodology outlined by Kuechler and Vaishnavi (2008). The prototype was implemented into an online learning platform to motivate girls to engage in computer science. Finally, we evaluated whether the prototype met the DPs with a seventh-grade class at a German school.

Cues for a Gender-Sensitive Conversational Agent Design

CAs, also known as digital assistants, or chatbots (Maedche et al., 2019), are widely used in different domains, including education, healthcare, and business (Maroengsit et al., 2019). In education, CAs offer students the opportunity to ask for help without having to be assisted by real teachers. Additionally, the careful incorporation of design cues in CAs can transform them into inclusive design interventions on online learning platforms. This approach aligns with Kizilcec and Saltarelli's (2019) principles of psychologically inclusive design, aimed at engaging underrepresented groups. Feine, Gnewuch, et al. (2019) identified four social cues to improve human likeness in CAs: verbal, visual, auditory, and invisible. These cues include, e.g., small talk, gender, ethnicity, age, gestures, and facial expressions of the CA. Properly selecting these social cues is crucial. While the right choices can increase user engagement and make CAs appear more relatable, poorly chosen cues can distract users' attention (Feine, Morana et al., 2019), e.g. causing them to focus on the CA's appearance rather than content. In our study, we focus primarily on gender cues and how the design of text-based CAs can be applied to support girls in online computer science courses. In their study, Zabel and Otto (2021) discussed how the match and mismatch of designer and user influences the user's perception of the CA without explicitly using gender cues. Gender match had a significant positive effect, while age match and the use of gender-sensitive language had no effect. In contrast, Schlimbach et al. (2023) found a positive impact on competence, trust, credibility, and empathy when using gender-sensitive language in CAs. They also explored differences in visual representation, using a male or female chatbot with or without a physical disability. Showing a physical disability significantly increased trust, credibility, and empathy, while decreasing competence and dominance. Using the female avatar instead of the male one significantly decreased competence, trust, credibility, and dominance. A study with 128 participants by McDonnell and Baxter (2019) further examined if the gender of the CA affects users' overall satisfaction. The results showed that the gender of a CA affects users' overall satisfaction and perceptions of gender stereotypes. Users preferred a CA with a male appearance for stereotypical male topics like mechanics. Related to this outcome, previous research found that gender stereotypes are applied based on the CA's voice (Nass & Moon, 2000) or appearance, where negative stereotypes were more often attributed to female representing chatbots (Brahnam & De Angeli, 2012). Forlizzi et al. (2007) show that users preferred CAs designed in a style that matched the stereotypical gender assignment in the domain, such as a male chatbot for coding problems. Arguing for gender as a primary design feature. Based on findings like these, McDonnell and Baxter (2019) introduced an ethical debate about whether biases and stereotypes should be deliberately reinforced, such as using a male CA for technical advice, to increase user trust. Fossa and Sucameli (2022) argue for a balance between efficiency and social acceptability to mitigate stereotypes in the long run without compromising utility. This solution raises the problem of determining which cases of bias are acceptable and which are not and who should judge them. Aside from the ethical issue of gender

cues in CAs, contradictory results from previous studies also raise questions about the effectiveness of gender cues. Arroyo et al. (2009) and Krämer et al. (2016) present that CAs of the opposite gender can be particularly effective in learning depending on the subject. Krämer et al. (2016) investigated the effect of gender cues in chatbots on mathematical tasks. Their results showed that participants performed better and exerted more effort when interacting with an agent of the opposite gender that shows rapport. However, the authors noted that the virtual agent in their study did not act as a competent role model, as described in the literature, but rather as a motivating interviewer. Therefore, they conclude that the mere presence of a woman in a learning scenario does not automatically increase motivation and learning. However, this does not rule out the possibility that the results would differ if a competent role model was provided. Arroyo et al. (2009) argue that female students might prefer the male agent because they project their stereotypes to it. In contrast, Rosenberg-Kima et al. (2010) found that individuals, particularly women, often preferred CAs of their own gender, perceiving them as more relatable and inspiring. Similarly, a study by Arroyo et al. (2011) contributes to this discourse by examining the impact of gender-matched animated agents in mathematics tutoring. Their findings reveal that female students benefited from interactions with female learning companions, reporting more positive learning experiences, reduced frustration, and increased confidence. This underscores the importance of considering gender cues in CA design to support underrepresented groups effectively. In addition, Ozogul et al. (2013) examined the influence of gender-matching versus gender-mismatching on students' learning outcomes and perceptions, following the similarity attraction hypothesis. Consistent with this hypothesis, female students viewed the classroom experience more favorably when the pedagogical agent matched their gender. However, contrary to expectations, male students rated the experience slightly higher when the pedagogical agent did not match their gender than when it did. The results of previous research are inconsistent and motivate for further research into the gender-sensitive design of CAs. In summary, when discussing gender cues, all studies summarized above utilize cues from the verbal or visual categories of the taxonomy of social cues (Feine, Gnewuch, et al., 2019). By altering the visual characteristics of the CA, including attributes such as gender, users' perceptions of the CA improved positively. (Arroyo et al., 2011; Rosenberg-Kima et al., 2010). Gender-sensitive language has been shown to positively influence perceptions of competence, trust, and credibility. The choice of words and the style of communication can reinforce or mitigate gender stereotypes, impacting user engagement and satisfaction (Schlimbach et al., 2023).

Multi-methodical Approach

To gain design knowledge on implementing gender-sensitive CAs, we applied the design science research paradigm by Hevner et al. (2008). We iteratively conducted four design cycles, as shown in Figure 1, following the methodology outlined by Kuechler and Vaishnavi (2008). We concluded each design cycle with an evaluation, the results of which were considered when implementing the subsequent prototype. We aim for relevant and rigorous artifact development through continuous design and evaluation cycles in constant exchange with the knowledge base and the environment.

Our methodological approach began in August 2022 with a literature review and workshops with didactic experts to identify relevant requirements for a CA that supports girls in online computer science courses. Following vom Brocke et al. (2015), we conducted a sequential literature search, aiming to find the relevant articles at the beginning of the review. To achieve this, we scanned bibliographic databases with keyword search to extract representative works. We developed the search string iteratively, screening the results of each iteration to see if the articles were fitting, the final search string used was:

("intelligent tutoring system" OR "digital mentor" OR "digital tutor") AND (chatbot OR "conversational agent" OR "dialogue system") AND ("e-learning" OR "online learning" OR "digital education" OR "online course" OR "digital learning") AND (coding OR "computer science" OR IT) AND (women OR girl OR female)

We conducted a comprehensive search for articles across commonly used academic databases, retrieving a total of 254 articles from ACM (7), AISel (24), Emerald Insight (8), ScienceDirect (8), SpringerLink (206), and Wiley (1). Despite also querying JSTOR, IEEE, Scopus, and Web of Science, no additional relevant articles were found in these databases. Subsequently, an additional 120 articles were extracted by screening Google Scholar, continuing until the articles presented were no longer relevant to the research topic. Out of these 374 articles, 316 were accessible. We removed duplicates, leaving 306 articles for further consideration, and screened the titles and tables of contents, excluding 184 articles. Of the 122 remaining

articles, we screened the abstract for relevance, leaving 28 articles. Finally, a full-text screening was performed on the remaining articles. This resulted in 13 articles from which we extracted requirements (Table 1). Additionally, we enriched requirements from the literature with requirements resulting directly from STT (Steele & Aronson, 1995). Lastly, we conducted an expert workshop aimed at gathering deeper insights. The workshop included ten participants, aged between 22 and 35, five experts on the gender-sensitive design of AI systems (four women, one man), and five on the gender-sensitive design of learning materials (two women, three men). These experts were specifically chosen for their different domain expertise, offering diverse perspectives, and experience with the design of gender-sensitive solutions. The workshop started with a 20-minute introduction to three user stories about girls wanting to learn computer science, followed by a 30-minute pair discussion where participants engaged in brainstorming and open dialogues to explore potential requirements for CAs to support these girls. Each pair consisted of one expert from each domain, promoting interdisciplinary perspectives. Subsequently, a 45-minute group discussion allowed each pair to present their findings, explaining the importance and relevance of the identified requirements. Other participants provided feedback, to foster a comprehensive evaluation of each point raised. After all presentations, the findings were summarized and aggregated, with an emphasis on capturing diverse viewpoints. We assumed that saturation was achieved as the latter presentations did not reveal any new requirements that hadn't already been discussed by previous groups. The discussions were held using an online whiteboard, which enabled real-time collaboration and visualization of ideas and was later summarized and documented by the researchers.

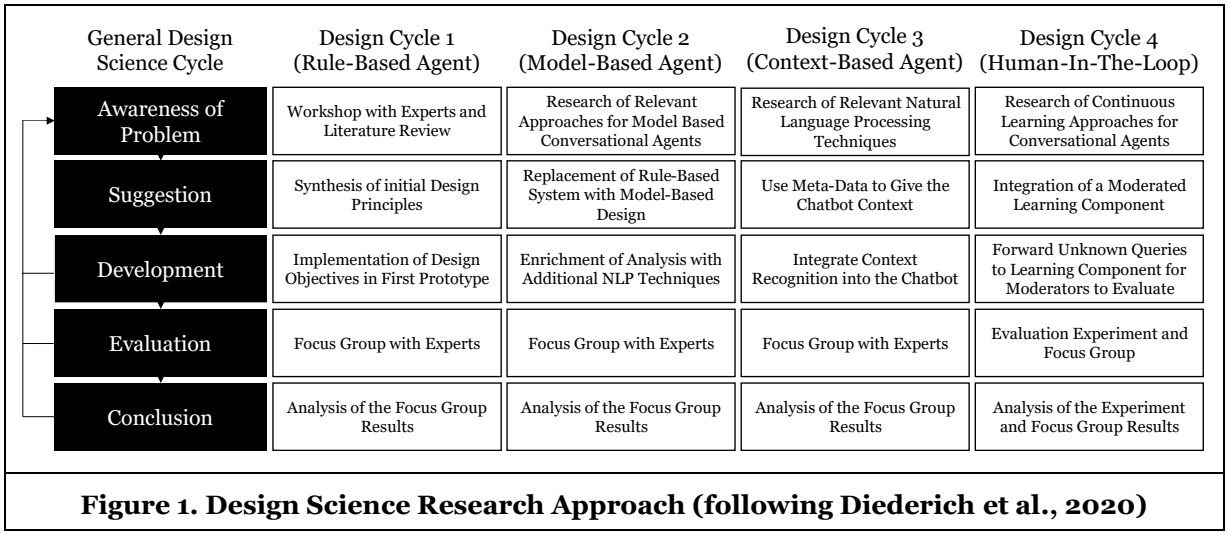

Figure 1. Design Science Research Approach (following Diederich et al., 2020)

The workshop and literature review resulted in 14 requirements; from which, we derived eight meta-requirements (MRs) and three DPs following the anatomy of Gregor et al. (2020) and the structure of Meier et al. (2021). The prototype was then implemented as a rule-based expert system and made publicly available on an online learning platform. In the subsequent evaluation phase, the prototype underwent testing by the same experts who participated in the initial workshop in focus groups. They reviewed the CA, discussed problems and good practices, and developed implementation goals with the researchers for the next iteration. In this iteration, they pinpointed the overly static nature of the CA as a significant limitation. This led to a necessary shift from a rule-based to a model-driven system, as the rule-based model could only respond based on specific keywords or phrases, limiting its adaptability. In contrast, a model-driven approach can predict the underlying meaning of inputs probabilistically, offering greater flexibility in responses. Additionally, the vast number of courses to be supported made it impractical to establish a comprehensive set of rules. A model-based system conversely requires only a few examples per course and question, making it more scalable and effective. While this shift represented a substantial technical modification in the implementation of the CA, it did not affect the integrity of the DPs. The second phase realized the change to a model-based approach to improve flexibility and scalability. The updated prototype, developed with Rasa Open Source, was re-evaluated by the focus group, prompting the realization that the CA should prioritize understanding the context of the learning materials to enhance responsiveness. This change was necessary because broad questions like ‘What do I need to do?’ require different answers based

on the context. The third cycle implemented the prototype using the Rasa action server to allow the CA to respond to questions context-sensitively. We configured the CA to react to the meta-data of user queries and adapt the responses accordingly. The new prototype was re-evaluated, which resulted in the final cycle, focusing on the need to improve the expandability of the CA for technically inexperienced educators, as the raw training files of the CA model were too complex to extend. Standardized solutions were initially explored but ultimately deemed too complex for these users. As a result, we extended the Rasa architecture with a self-developed moderated learning component (MLC). The training data for this prototype was generated using the training data and the MLC with the input data of previous cycles. Finally, we tested whether the prototype met the DPs in a summative evaluation with a seventh-grade class. Students tested the CA on an online learning platform and provided feedback through a survey and focus groups.

Table 1. Concept Matrix of Requirements

Requirements			Albornoz-De Luise et al. (2022)	Faenza et al. (2021)	Jung et al. (2020)	Kuhail et al. (2022)	Kuttal et al. (2021)	Latham (2022)	Lobo et al. (2022)	Martha and Santoso (2019)	Oliveira et al. (2021)	Scholten et al. (2017)	Wellnhammer et al. (2020)	Yildiz and Rızvanoğlu (2021)	Zhang and Aslan (2021)	Steele and Aronson (1995)	Expert Workshops
Promote Stereotype-Neutral Behavior	R1	5		x			x						x			x	x
Encourage Engagement	R2	7		x		x		x		x	x		x			x	
Emotional Intelligence	R3	10	x		x	x	x	x		x		x		x	x	x	
Non-Intrusive Design	R4	3			x			x				x					
Adaptation to the Individual Needs of Users	R5	13	x	x	x	x	x	x	x	x	x		x	x	x		x
Visual Adaptability and Appeal	R6	4								x		x	x				x
Trustworthiness	R7	6					x	x				x	x	x			x
Acceptance	R8	3				x							x				x
Consideration of Social Context	R9	10		x	x	x			x	x	x	x	x	x		x	
Sense of Belonging	R10	3		x						x	x						
Support the Building of Confidence	R11	3		x												x	x
Alleviate Anxiety	R12	2								x						x	
Mitigate Performance Pressure	R13	2					x									x	
Alleviate Cognitive Overload	R14	5			x		x						x			x	x

Design Principles for Gender-Sensitive Conversational Agents

Girls face unique challenges when engaging in online computer science courses due to gender stereotypes. To address these issues, promoting *engagement through gender sensitivity* is crucial (MR1). As Jung et al. (2020) noted, gender stereotyping in CA interactions can significantly impact how information is received and valued by users, especially in the computer science domain. Faenza et al. (2021) highlight the persistence of the digital gender gap, attributing it significantly to entrenched stereotypes that paint the computer science sector as dominantly masculine and unwelcoming to women, thereby deterring their participation. Therefore, designing inclusive CAs, not only in their gender representation but also in their interaction style, can mitigate the negative effects of stereotyping. Further, providing *Intuitive and Empathic Responses* (MR2) plays a crucial role. Scholten et al. (2017) highlight the positive impact of CAs that can detect and empathically respond to user frustration. Especially for underrepresented groups, such as women in computer science, it is essential to mitigate user frustration due to high dropout rates. Albornoz-De Luise et al. (2022) underscore the importance of detecting cognitive states like concentration and frustration to offer effective support. Such capabilities ensure the CA can offer real-time, context-sensitive support, making the learning experience more intuitive and aligned with the user's emotional needs. Therefore, the requirements for gender-sensitive engagement and intuitive and empathic responses in a CA design lead to the following DP (Figure 2):

DP1: *Engineer an emotionally intelligent and gender-sensitive interface within the conversational agent to encourage contextually relevant and engaging interactions for girls in online computer science courses.*

The *Personalization of Experience and Visual Interaction* (MR3) is essential for effectively engaging girls in online computer science courses. As highlighted by Göritz et al. (2022) and Krämer et al. (2016). CAs need to consider individual users' needs, goals, and social contexts, enabling tailored interactions with each student, e.g., by adapting the choice of words or the visual representation to the different realities of the users. This adaptability ensures that interactions are relevant and resonate with the learner's personal journey. Kumi–Yeboah et al. (2017) emphasize that culturally relevant teaching can increase minority engagement, making it easier for them to align course content with their individual realities and identities. This is particularly important in counter-stereotypical fields such as computer science. The research of Oliveira et al. (2021) on tailoring content presentation based on students' personality traits and contexts underscores the importance of personalization in facilitating effective learning engagements. Furthermore, the ability of a CA to exhibit live emotion can enhance the interaction by making it more engaging and personalized (Jung et al., 2020). *Establishing Credible and Relatable Interactions* (MR4) is crucial to a supportive learning environment. As Jung et al. (2020) mentioned, imitating human-like traits in chatbots can foster more authentic interactions and build a trusting relationship, which is crucial for engaging learners effectively. As Latham (2022) points out, trust in the CA is vital for mitigating negative feelings such as frustration and enhancing the learning experience. The credibility of a CA, influenced by content quality and speech parameters (Wellnhammer et al., 2020), alongside trust in the agent's functionality (Khosrawi-Rad et al., 2022; Morana et al., 2020), is a key factor for the acceptance and effective use of innovative educational solutions. This aspect is particularly vital for girls, who may require additional affirmation of the CA's reliability to feel secure engaging in the male-dominated field of computer science. *Cultivating an Inclusive Community* (MR5) is further important for supporting girls in computer science. The awareness of social context, as Lobo et al. (2022) suggest, enables CAs to manage interactions effectively within a broader social and AI ecosystem. Initiatives like the "Digital Girls" summer camp, highlighted by Faenza et al. (2021), provide exclusive support to female students, and promote an inclusive community by exposing them to computer science in a supportive environment. These efforts are crucial for girls who may feel isolated or marginalized in predominantly male environments. Therefore, considering the importance of a personalized experience and visual interaction, the establishment of credible and relatable interactions, and the cultivation of an inclusive community for girls in computer science courses, we propose the following DP (Figure 2):

DP2: *Develop a trust-centric and customizable learning environment through the conversational agent that fosters a sense of community among users to support personalized and empowering educational experiences for girls in online computer science courses.*

Enhancing User Empowerment (MR6) plays a crucial role. As outlined by Jung et al. (2020) and Karra and Lasfar (2021), CAs can offer tailored support and feedback like a real teacher, fostering student empowerment. This empowerment is further enhanced by their ability to provide instant corrective feedback, generate automatic scoring, and assist with revisions throughout the learning process, as discussed by Zhang and Aslan (2021). Moreover, by leveraging students' personality traits and context, as mentioned by Oliveira et al. (2021), CAs can present content in a manner that resonates with each student's unique learning style, further empowering them by making the learning experience more relevant and engaging. This empowerment is particularly significant for girls in computer science, helping them overcome the unique challenges and barriers they face in a field where they are underrepresented. The *Simplification of Information Processing* (MR7) is another critical aspect of breaking down psychological barriers and thereby reducing uncertainty among underrepresented groups. Agent-based learning technologies offer the opportunity to provide low-threshold assistance to insecure students who may be reluctant to ask for help due to a lack of self-efficacy in the computer science context (Leider & Strobel, 2020; Rosenberg-Kima et al., 2010). As Albornoz-De Luise et al. (2022) suggested, the transition to a CA offers a more natural interface that facilitates easier interactions. This simplification is crucial for girls in computer science who might otherwise feel overwhelmed by technical complexities or hesitant to participate in discussions in male-dominated settings. Furthermore, Latham (2022) emphasizes the importance of carefully designed tutoring conversations to mimic a human-like learning experience involving expert knowledge of pedagogy and the domain being taught. Oliveira et al. (2021) also highlight the significance of providing adaptive and personalized learning experiences through natural language interactions, which simplifies how information is processed by students, making learning more efficient and less overwhelming. Considering these challenges and requirements, we propose the following DP (Figure 2):

DP3: *Facilitate empowering and streamlined interactions with the conversational agent through simplified dialogues to bolster comprehension and foster independence for girls in computer science courses.*

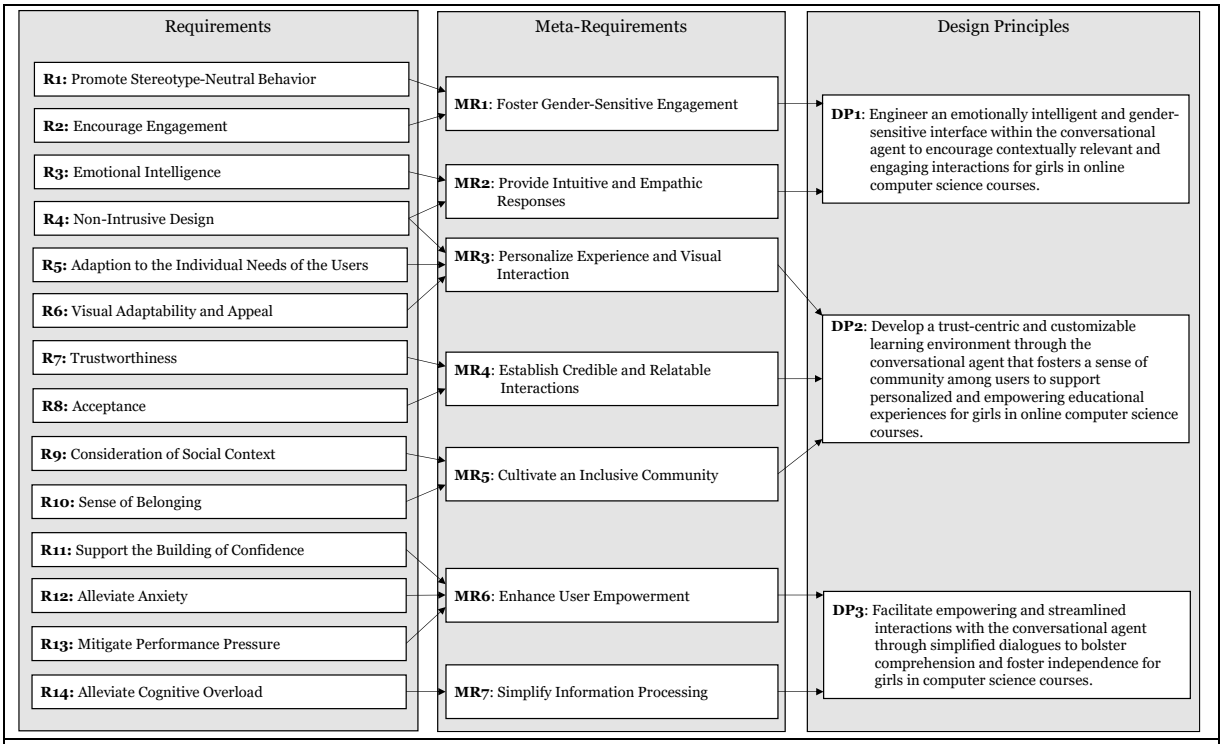

Figure 2. Synthetization of Meta-Requirements and Design Principles

Instantiation of the Design Principles

To instantiate the DPs, we implemented a technical infrastructure designed to realize the DPs and create an engaging, gender-sensitive, trust-centric, and empowering learning environment within an online platform for computer science education targeted at girls. The CA is a pivotal component of this platform, offering various courses in the computer science domain. Figure 3 illustrates the layered system architecture, demonstrating how each component contributes to realizing our DPs. Few components are not directly responsible for the instantiation of the DPs but provide a framework for the CA as a whole: The agent, as the core of the system, manages communication with all the other components and the user and loads the model, which predicts the intent of the user's input and provides an output that the other components manipulate. The policies, which manage the general flow of conversations, and the database, which stores the files, e.g., the model, the training data from which the model is created, and the previous conversation data in the tracker store.

To instantiate DP1 the CA's emotionally intelligent and gender-sensitive interface is engineered to discern the nuances of the user's input through a natural language understanding pipeline. This allows the agent to comprehend the intent behind a query and perceive the subtle cues, facilitating a contextually relevant and supportive interaction. The architecture supports this by considering previous interactions with the user from the log store, enabling the CA to adapt its communication style and thus providing a tailored and engaging experience. The agent's design features a female appearance slightly older than the target group, further enhancing relatability and aspiration, serving as a role model, and strengthening the emotional connection with the user.

To realize DP2, trust is cultivated through tailored textual responses, as reflected by the agent's interaction with the log store, to ensure continuity and personalized attention. The CA's learning environment extends personalization through the recommender system, which leverages user preferences and behavior to offer personalized course suggestions. While the data for the personalization of responses is provided from the

log store and recommender, the action server processes the data and performs the adaptation. Further, the system's visual customization is realized in the CA's design, allowing users to adjust visual elements such as clothing, hair color, and skin color, fostering comfort and identification. These features are integral to creating a learning environment that is personalized to each girl's academic needs and her personal and cultural identity, laying a foundation of trust and security. Additionally, the MLC allows moderators to oversee the learning content, ensuring that the CA's responses remain accurate and appropriate, reinforcing the environment's trustworthiness.

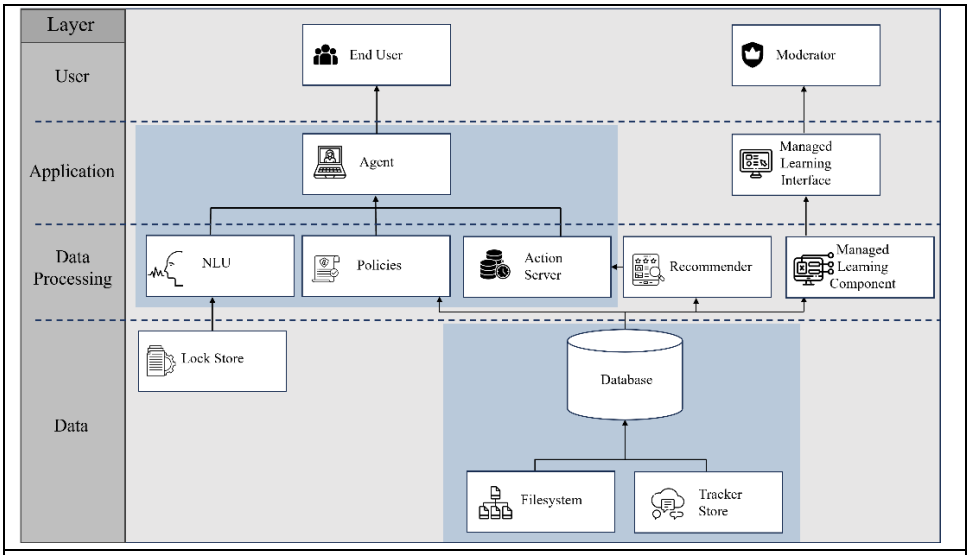

Figure 3. Layer Architecture of the Conversational Agent

To empower the girls and streamline interactions, and thereby implement DP3, the CA was trained on a dataset created specifically for the target group, emphasizing clarity and ease of use, thereby supporting the girls' comprehension, and promoting independence. This is achieved by the MLC, enabling non-technical moderators to manage and refine the training data that drives the CA's responses. The system ensures that the conversations remain relevant and supportive of independent learning by allowing the moderators to update the dialogue based on real interactions. The tracker store is integral to this process, offering a repository of previous conversations which enable continuous dialogue quality improvement. Deciding against a self-learning system mitigates the risk of compromised dialogue quality due to unsupervised inputs, thereby sustaining a high standard of didactic interaction that is critical for effective learning.

Evaluation Experiment

The evaluations we conducted throughout the research process can be categorized in the framework for evaluation in design science by Venable et al. (2016), which provides a two-dimensional characterization of evaluations. The framework classifies, on the one hand, according to the evaluation's purpose (formative or summative) and, on the other hand, according to the evaluation's paradigm (artificial or naturalistic). As described previously, the evaluations of the first three design cycles aimed to quickly and continuously derive and directly implement improvements to the CA. In contrast, the final evaluation as part of the fourth design cycle aimed at a summative evaluation. From the summative evaluation results, we derive generalizable findings that provide a basis for creating shared meanings of CAs. To achieve high internal validity of our results, we conducted them in a naturalistic setting with students in school. As part of the fourth design cycle, we conducted the summative evaluation of the CA with a German school class. We specifically chose a comprehensive school for the evaluation to get input from students with different levels of academic performance. Seven researchers from information systems, cognitive science, and psychology participated as instructors, moderators, and observers in the evaluation process, which was carried out in three phases. In the first phase, 16 girls and 11 boys aged 12 to 13 tested the CA. We selected this age group because their development stage is marked by a significant increase in stereotype awareness, making them ideal for this evaluation (McKown & Weinstein, 2003). To facilitate quantitative comparisons, we

conducted surveys with both boys and girls. We instructed them to interact with the CA while participating in computer science courses on an online learning platform for 30 minutes. During this phase, the students got a first impression of the CA, and the researchers observed the students interacting with the CA. To contribute to the research question, in the second phase, we quantitatively surveyed (Table 2) the fulfillment of our derived DPs.

Table 2. Scales Used for Evaluation

Construct	Established scale	Example-Item (translated)	Cronbach's α
Anthropomorphism (DP1)	Nowak et al. (1999)	I perceive the Agent as a person.	0.96
Attractivity (DP1)	Kemper et al. (2012)	I find the Agent visually appealing.	One Item Scale
Identification (DP1)	Basil (1996)	I can identify well with the Agent.	0.74
Recognition and facilitation of users' goal and intent (DP2)	Borsci et al. (2022)	I find that the Agent understands what I want and helps me achieve my goal.	0.71
Trustworthiness (DP2)	Borsci et al. (2022)	I had the feeling that I could trust the Agent.	0.9
Perceived Usefulness (DP3)	Davis (1989)	With the Agent, I can learn to program faster.	0.8
Accessibility (DP3)	Borsci et al. (2022)	It was easy to find the Agent.	0.7

DP1 aims to design an emotionally intelligent and gender-sensitive interface that ensures engaging and contextually relevant interactions to meet users' needs. DP1 was evaluated by measuring anthropomorphism using the scale developed by Nowak et al. (1999), attractiveness using a single-item scale by Kemper et al. (2012), and identification using Basil's (1996) scale. These measures assess the CA's ability to simulate human-like interactions, be visually appealing, and evoke a sense of familiarity. DP2 implies the provision of a trust-centric and customizable CA intervention. Accordingly, we operationalized this by incorporating subscales of the Usability Scale developed by Borsci et al. (2022). These subscales assess the recognition and facilitation of users' goals and intentions, as well as trust. For DP3, we measured the facilitation of empowering and optimized interactions with the CA using the Accessibility subscale of the Chatbot Usability Scale by Borsci et al. (2022) and the Utility-scale by Davis (1989). We translated the English scales into German, adapted the wording where necessary so that they were easy for children to understand, and then had them cross-checked by six researchers. The students answered on a seven-point Likert scale ("1 - strongly disagree" to "7 - strongly agree"). The scales we measured with more than two items have a Cronbach's alpha higher than .70, indicating good internal consistency (Blanz, 2021). In the third phase of the evaluation process, we conducted two focus groups of six girls to gain insights into the instantiation of the DPs. These findings can be added to the knowledge base for the scientific community. We decided to include only girls in the focus groups to create a comfortable and safe environment where girls could openly share their perceptions of the CA. The focus groups lasted 30 minutes and followed the guidelines of Myers and Newman (2007). A female researcher moderated the focus groups to provide the girls with comfort and minimize social dissonance. In addition, two observers attended each focus group. They took notes to replace an audio recording of the interview to give the girls the feeling that they could speak openly. A semi-structured guideline allowed us to respond individually to the girls' suggestions. The first part of the focus group guideline aimed to give the girls a sense of trust by thanking them for their help so far, asking them to share their experiences with the CA, and emphasizing that there are no wrong answers but that all feedback is essential and helpful. Next, the focus was further discussing suggestions, comments, and criticisms about the CA. In follow-up workshops, the six researchers shared their experiences based on their observations and notes collected and developed four areas for further CA development.

The evaluation results are divided into quantitative measurement of fulfillment and qualitative exploration of areas for further CA development. The average fulfillment is depicted in Figure 4, separately for 11 boys and 16 girls. The study's primary goal was to fulfill the DPs from the girls' perspective. We measured the DPs from the boy's perspective to evaluate if they rejected the CA; meeting the DPs from their perspective would serve as a bonus. Figure 4 shows that, on average, the girls agreed with the design requirements we

measured quantitatively, except for the anthropomorphism scale. However, the boys disagreed with three of the eight measurements (recognition and facilitation of users' goal and intent, identification, and anthropomorphism). To test gender differences for statistical significance, we calculated T-tests for independent samples. First, we tested the sample for normal distribution and homoscedasticity. These sampling assumptions were not met for four of the eight measurements (perceived usefulness, attractivity, accessibility, and trustworthiness); thus, we calculated the nonparametric alternative, the Wilcoxon rank sum test. Gender differences were significant for three of the eight measurements. Girls rated perceived usefulness higher (Mdn = 5.5) than boys (Mdn = 4.67), $z = -1.48$, $p = .01$, $r = -.55$. Girls agreed more ($M = 4.88$) than boys ($M = 3.64$) with the recognition and facilitation of users' goal and intent, $t(25) = 2.74$, $p = .01$, $d = 1.07$. Furthermore, girls rated the CA higher in attractivity (Mdn = 6) than boys (Mdn = 4), $z = -1.80$, $p = .03$, $r = -.42$.

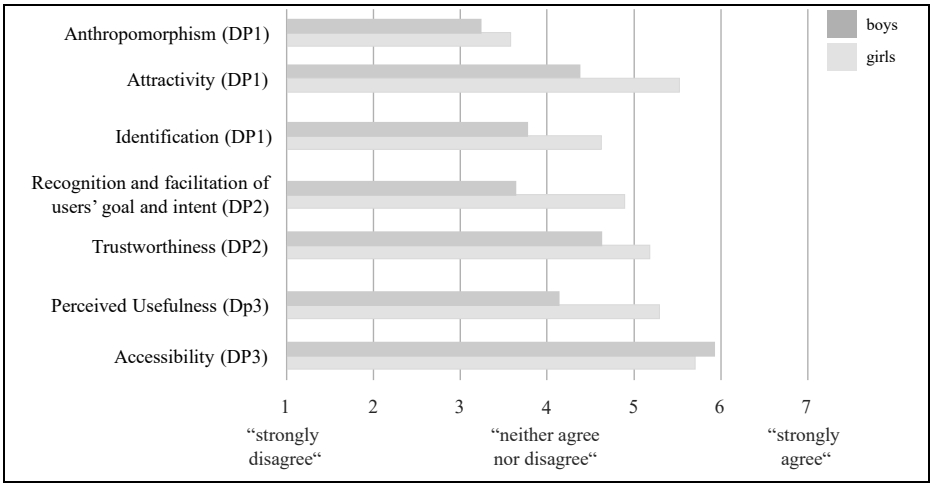

Figure 4. Average Agreement with the Evaluated Constructs

We derived four areas for further CA development from the focus groups. These include (1) expanding its word base, (2) adopting a more natural communication style, (3) proactively reaching out to the users, and (4) increasing its visibility on the platform. On the one hand, expanding the word base enables more technical questions that refer, for example, to the course offerings. On the other hand, the students wanted more personal questions to be answered. This includes answers to simple small talk questions and more private questions about "girl problems". Students discussed that asking a CA more personal questions that require more trust might be easier because they don't have to fear judgment. Furthermore, they criticized the CA for always replying with the same phrase when not having an answer. The students wished for more variability and the possibility to ask other users for an answer via chat. The students made several small suggestions to achieve a more natural communication style. Like other messenger apps, the chat should never reset itself. Furthermore, depending on the length of the text, the reply should be delayed by a "Is typing...". The students also noted that the CA does not ask questions back, as in most conversations between people. They suggested that the CA could learn more about the user's name, age, and coding experience by asking questions. Moreover, the students wanted the CA to use more informal language, shorter response texts, and more emoticons to communicate more like a peer. In addition, students wished the CA would be more proactive in engaging with them. For example, the CA could welcome users when they enter the platform and introduce itself to them. It could also check in with them occasionally and ask if they need help. Students suggested that the CA could proactively give them personalized course recommendations, for example, after completing a course. The fourth and least mentioned area for further development of the CA was to improve its visibility on the platform. There were a few suggestions from students to place the CA more in the center and to widen the chat window. In addition to areas identified for further development, the feedback from the focus groups also highlighted several strengths of the CA. The girls responded positively to the CA's immediate and contextually relevant answers, appreciating its capacity to assist efficiently in answering questions and providing information. They enjoyed the potential for customization, such as the ability to change its appearance, enhancing the personal connection with the CA. Furthermore, the girls expressed a preference for the CA's female appearance, noting that they would

not have wanted a male one. The overall interaction with the CA was perceived as engaging and supportive, making the learning experience more enjoyable. This feedback underscores the CA's effectiveness in facilitating an engaging and helpful educational intervention.

Discussion

With our study, we have developed an approach to motivate and support girls in computer science with a CA. Previous research has shown that role models are a well-established intervention in motivating underrepresented groups to engage with a topic they may not initially identify with (Rosenberg-Kima et al., 2010). For instance, providing teachers with different characteristics on an online learning platform makes it possible to address different types of learners and provide room for identification. However, employing various teachers with varying characteristics for the same course to address all learners is hardly feasible and uneconomical. As CAs can also act as role models, their ability to adapt to the needs of learners and fulfill the required characteristics is of great importance. A CA can be personalized automatically through AI algorithms or manually with just a few clicks to meet the required characteristics. Providing students with personalized learning content, processes, and environments has been a long-standing educational challenge. Personalization and differentiation are essential in implementing didactic principles attentively, especially for underrepresented students who may not initially identify with the subject (Brooks et al., 2018; Kizilcec & Kambhampaty, 2020). Differentiation makes it possible to address different needs within the target groups. The possibility to personalize a CA on an online learning platform can therefore immediately impact the experience of many students worldwide.

The overall results of our evaluation indicate that our instantiation of the DPs support the girls to identify and engage with computer science. While our finding that girls prefer gender-inclusive design may seem obvious, this is not always true. As highlighted in the conceptual review by (Stumpf et al., 2020), historical research in Human-Computer Interaction has often focused on males, resulting in IT artifacts, such as learning platforms, that are predominantly and unintentionally tailored for male users. Consequently, contemporary Human-Computer Interaction research should strive to redress this imbalance and promote diverse design choices for gender-responsive IT artifacts. This initiative is imperative to mitigate inherent biases in technology and, consequently, to promote greater gender equity in IS.

The focus groups extended the knowledge generated by the survey and emphasized the relevance of the DPs. Participants appreciated the immediate, context-sensitive responses from the CA, emphasizing the importance of personalization in creating a supportive learning environment. Feedback suggesting improvements, such as incorporating emoticons, small talk, and a more casual dialogue style, underscored the desire for a more human-like interaction, aligning with the principle of engineering emotionally intelligent interfaces to enhance user engagement. One way of achieving an improved and, therefore, more natural communication style for the CA would be to use Large Language Models (LLMs), which can respond naturally to user input if the role assignment is correct. We decided against an LLM because precise LLMs were not available in an open-source version when we began the implementation and due to privacy concerns with commercial solutions since learning data is highly sensible. Furthermore, inaccurate, or highly stereotypical LLM results could harm a culturally relevant and gender-sensitive teaching style. To minimize the risks but still profit from LLMs in future research, we propose the introduction of a supervising layer. This layer would check that the LLM does not give stereotypical or offensive answers.

The results of the focus groups suggest that a CA incorporating our DPs can enhance the attraction of counter-stereotypical subjects, such as computer science, to girls and thus motivate their engagement. Our study aimed to mitigate stereotype threat resulting from societal expectations, aligning with the call from Kizilcec and Saltarelli (2019) to investigate how different cues influence user interaction based on context and identity, thereby enriching our understanding of STT. For girls, the challenge of reconciling learning content with their gender identity creates a state of psychological tension that often leads to anxiety. This anxiety may deter girls from situations in which they fear embarrassment, such as the fear of being judged by male teachers in computer science classes. As a result, they may avoid reinforcing negative stereotypes about girls in the field. Due to a perceived lack of self-efficacy, girls may develop insecurities and be reluctant to seek help from teachers. Previous research has suggested that online learning platforms should actively provide accessible help channels to address this psychological barrier (Leider & Strobel, 2020). Additionally, anonymously navigating an online learning environment is critical in this context (Leider & Strobel, 2020). In our study, these low-threshold and anonymous help channels are facilitated by our DPs

developed through the design science research methodology. As the evaluation shows, the implemented CA can address these issues and support the girls, thus empowering them in computer science education.

Our research contributes to a deeper understanding of DPs' relevance to CAs. While our research focuses on supporting girls in online computer science courses, we assume high generalizability to other underrepresented groups in different domains. We extend the design knowledge for the IS discipline using a social sciences theory. Thus, we emphasize the need for an interdisciplinary research approach to key IS topics, e.g., the design of a CA as a technical artifact, by incorporating insights and expertise from other disciplines such as psychology, social sciences, and education. Despite its relevance, STT has rarely been used in IS research as a kernel theory for gender-sensitive design. It has only been addressed in a few IS journals, such as Computers in Human Behavior. However, inclusion is rapidly gaining prominence in the IS community, as evidenced by calls for papers on social inclusion and the social and ethical implications of information and communication technology use at IS conferences such as ICIS24, AMCIS24, ECIS24, and PACIS24. Our study contributes to research in the field of social inclusion by providing insights on how to implement design cues in CAs to support in a gender-sensitive matter. Our research contributes to the ongoing discussion on CA design. Consistent with the findings of Arroyo et al. (2011) and Rosenberg-Kima et al. (2010), our results reveal a preference for CAs that match the user's gender, as evidenced by the girl's specific disapproval of male agents in the focus groups. This contradicts the studies by Arroyo et al. (2009) and Krämer et al. (2016), which suggested a preference for agents of the opposite gender. While our findings cannot affirm Ozogul et al. (2013), who reported a preference for female agents among both boys and girls, they do indicate that boys were not rejecting the agent based on its gender. Furthermore, our study diverges from McDonnell and Baxter (2019), who observed gender stereotyping of the agent's competence; conversely, we found no evidence of such stereotyping among the girls in our study. Our proposed DPs incorporate gender cues and show that implementing a CA as a gender-sensitive design intervention is not only about its visual appearance but also about its demeanor, such as the understanding of emotions, and technical features, such as the adaptability of responses. Based on the taxonomy by Feine and Gnewuch et al. (2019), our research therefore extends previous findings by incorporating not only verbal and visual social cues but also invisible cues such as the adaptability of responses, which were highly valued by the girls in our focus groups and should be expanded upon. This suggests that for text-based agents, integrating cues from all three categories is essential for effective design. Building on the research avenues of Diederich et al. (2022), our work demonstrates that CAs tailored to specific groups in a domain can mitigate gender stereotypes and enhance engagement with the domain. This research addresses the research avenues "User Characteristics and Adaptive CA Designs" and "Ethical Implications of Designing and Interacting with CAs".

In addition, our research provides a concrete, practical contribution by deriving clear guidelines that practitioners can apply to make their CAs more gender-sensitive. Learning platform providers can improve the quality of their platform for girls and potentially other underrepresented groups by integrating a customized CA, thereby retaining them on the platform. As shown in the evaluation, this can be achieved without compromising the quality of the learning materials for the previously overrepresented group if the design aspects are chosen carefully. Expanding the target group could increase the number of learning platform users, potentially yielding monetary benefits for the platform providers.

Our research findings also have significant societal relevance. By advocating for a more gender-sensitive design of online learning platforms, we address the inequality issues in education raised by the European Commission (2020) and increase the pool of potential students who would otherwise risk not exploring subjects like computer science due to stereotypical perceptions, we thereby address societal challenges such as skills shortages. The underrepresentation of women in IT is a well-documented issue influenced by various social, cultural, institutional, and individual factors that interact with and perpetuate one another (Botella et al., 2019). Notably, research suggests that innate differences in ability do not account for this gap. Instead, implicit, and explicit biases prevalent in the tech industry contribute significantly to shaping perceptions of value and the environment within the field (Charlesworth & Banaji, 2019). Engaging an underrepresented group with a platform carefully designed interventions, such as CAs, can address these challenges appropriately. There needs to be more than the gender-sensitive design of the CA to achieve this goal; instead, the overall design of the platform is crucial. However, as the interface for user interaction, the CA is critical to this overall picture. The societal challenges addressed by online learning platforms with a CA tailored to an underrepresented group, in general, can be diverse, assuming that education and intergroup imbalance are the triggers for these issues.

Conclusion, Limitations, and Outlook

Based on our research question, we investigated how a CA should be designed to support girls in online computer science courses. We identified requirements from the literature and expert interviews and combined them with the STT. We derived seven MRs and three DPs. Finally, the DPs were instantiated as a CA on an online learning platform for computer science courses and subsequently evaluated. The application of our multidisciplinary approach, including elements from IS, social sciences, psychology, and education grounded in STT, has led to the development of DPs that foster a gender-sensitive online learning experience, empowering girls in computer science education. This is supported by our evaluation results, which underline the effectiveness of these DPs in addressing the specific needs of girls. As with all research efforts, our study has some limitations. The current results of our study demonstrate the collective evaluation of the DPs, not allowing us to infer the individual relevance of each principle. In future research, we intend to integrate and evaluate these DPs independently. In addition, the evaluation results have highlighted the need for a more accurate implementation of perceived anthropomorphism within the CA. In this regard, the integration of an LLM may be beneficial. Moreover, our feedback collection for the CA featured limited diversity, e.g. by predominantly involving European experts and students. Future studies should aim to include participants from more varied backgrounds to enrich the diversity of perspectives. Additionally, it's important to note that a causal relationship between an inclusively designed CA and girls' perceived insecurities and threats in online computer science courses cannot be definitively established due to the need for an experimental evaluation design with control groups. Future research should address this challenge to make a more robust contribution to STT. We call for continued evaluation of the DPs we have developed to further contribute to STT research. To measure the reduction in stereotype threat mitigated through a CA effectively, future research must compare students' learning outcomes when interacting with a CA that incorporates our DPs with those of a control group in the long run. When considering long-term studies, we should also observe if students learning with a gender-sensitive CA make different decisions in the future than students who learned without one, e.g., the choice of their study field or their job choice. Nevertheless, the overall positive feedback from the evaluation underlines the potential of technological interventions to design gender-sensitive online environments. The potential lies, for example, in increasing user reach on online platforms and in the socio-economic impact of greater inclusion in society. We recognize that the transferability of the DPs to other underrepresented groups requires further research to ensure its relevance. Our study contributes to the academic discourse and offers practical, actionable insights for designing technologies that foster equality and diversity in education and beyond. The prospect of utilizing our DPs across different domains heralds the potential of a more inclusive future, where technology serves as a bridge rather than a social barrier.

References

Abdelhalim, E., Anazodo, K. S., Gali, N., & Robson, K. (2024). A framework of diversity, equity, and inclusion safeguards for chatbots. *Business Horizons*. **https://doi.org/10.1016/j.bushor.2024.03.003**

Albornoz-De Luise, R. S., Arevalillo-Herráez, M., & Arnau, D. (2022). Using Open Source Technologies and Generalizable Procedures in Conversational and Affective Intelligent Tutoring Systems. *International Conference on Artificial Intelligence in Education*, 55–58. Cham: Springer International Publishing. **https://doi.org/10.1007/978-3-031-11647-6_9**

Arroyo, I., Woolf, B. P., Royer, J. M., & Tai, M. (2009). Affective gendered learning companions. *Frontiers in Artificial Intelligence and Applications*, *200*(1), 41–48. **https://doi.org/10.3233/978-1-60750-028-5-41**

Arroyo, I., Woolf, B. P., Royer, J. M., Tai, M., Muldner, K., Burleson, W., & Cooper, D. (2011). Gender matters: The impact of animated agents on students' affect, behavior and learning. *Technical report UM-CS-2010*.

Basil, M. D. (1996). Identification as a mediator of celebrity effects. *Journal of Broadcasting & Electronic Media*, *40*(4), 478–495. **https://doi.org/10.1080/08838159609364370**

Blanz, M. (2021). *Forschungsmethoden und Statistik für die Soziale Arbeit: Grundlagen und Anwendungen*. Kohlhammer Verlag.

Borsci, S., Malizia, A., Schmettow, M., Van Der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-

Based Conversational Agents. *Personal and Ubiquitous Computing, 26*, 95–119. **https://doi.org/10.1007/s00779-021-01582-9**

Botella, C., Rueda, S., López-Iñesta, E., & Marzal, P. (2019). Gender diversity in STEM disciplines: A multiple factor problem. *Entropy*, 21(1), 30. **https://doi.org/10.3390/e21010030**

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers, 24*(3), 139–153. **https://doi.org/10.1016/j.intcom.2012.05.001**

Brooks, C., Gardner, J., & Chen, K. (2018). How gender cues in educational video impact participation and retention. *13th International Conference of the Learning Sciences (ICLS) 2018, 3*.

Charlesworth, T. E. S., & Banaji, M. R. (2019). Gender in science, technology, engineering, and mathematics: Issues, causes, solutions. *Journal of Neuroscience*, *39*(37), 7228–7243.

Clayton, K. L., von Hellens, L. A., & Nielsen, S. H. (2009). Gender stereotypes prevail in ICT: A research review. *Proceedings of the Special Interest Group on Management Information System's 47th Annual Conference on Computer Personnel Research*, 153–158. **https://doi.org/10.1145/1542130.1542160**

Cocchi, A., Bosse, T., & van Pinxteren, M. (2023). Should Conversational Agents Care About Our Gender Identity? *International Workshop on Chatbot Research and Design*, 149–163.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly: Management Information Systems, 13*(3), 319–339. **https://doi.org/10.2307/249008**

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology, 56*(1), 5-18. **https://doi.org/10.1037/0022-3514.56.1.5**

Diederich, S., Brendel, A. B., and Kolbe, L. M. 2020. Designing Anthropomorphic Enterprise Conversational Agents. *Business and Information Systems Engineering, 62(3),* 193–209.

Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the Design of and Interaction with Conversational Agents: An Organizing and Assessing Review of Human-Computer Interaction Research. *Journal of the Association for Information Systems, 23*(1), 96–138. **https://aisel.aisnet.org/jais/vol23/iss1/9**

European Commission. (2020). Mitteilung der Kommission an das Europäische Parlament, den Rat, den Europäischen Wirtschafts- und Sozialausschuss und den Ausschuss der Regionen. Eine Union der Gleichheit: Strategie für die Gleichstellung der Geschlechter 2020-2025. **https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=COM%3A2020%3A152%3AFIN**

Faenza, F., Canali, C., & Carbonaro, A. (2021). ICT Extra-curricular Activities: The "Digital Girls" Case Study for the Development of Human Capital. *Springer Proceedings in Complexity*, 193–205.

Faik, I., Sengupta, A., & Deng, Y. (2024). Inclusion by Design: Requirements Elicitation with Digitally Marginalized Communities. *MIS Quarterly*, 48(1).**https://doi.org/10.25300/MISQ/2023/17225**

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., Morana, S., & Maedche, A. (2019) Designing a Chatbot Social Cue Configuration System. *ICIS 2019 Proceedings*. 2. **https://aisel.aisnet.org/icis2019/design_science/design_science/2**

Forlizzi, J., Zimmerman, J., Mancuso, V., & Kwak, S. (2007). How Interface Agents Affect Interaction Between Humans and Computers. *Proceedings of the 2007 Conference on Designing Pleasurable Products and Interfaces*, 209–221. **https://doi.org/10.1145/1314161.1314180**

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

Göritz, Lorena; Stattkus, Daniel; Beinke, Jan Heinrich; and Thomas, Oliver, "To Reduce Bias, You Must Identify It First! Towards Automated Gender Bias Detection" (2022). *ICIS 2022 Proceedings*. 10. **https://aisel.aisnet.org/icis2022/data_analytics/data_analytics/10**

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). Research perspectives: The anatomy of a design principle. *Journal of the Association for Information Systems, 21*(6), 1622–1652. **https://doi.org/10.17705/1jais.00649**

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2008). Design science in information systems research. *Management Information Systems Quarterly, 28*(1), 75-106.

Jung, H., Lee, J., & Park, C. (2020). Deriving Design Principles for Educational Chatbots from Empirical Studies on Human–Chatbot Interaction. *Journal of Digital Contents Society, 21*(3), 487–493. **https://doi.org/10.9728/dcs.2020.21.3.487**

Karra, R., & Lasfar, A. (2021). Enhancing education system with a Q&A Chatbot: a case based on open edX platform. *International Conference on Digital Technologies and Applications*. Cham: Springer International Publishing., 655–662. **https://doi.org/10.1007/978-3-030-73882-2_59**

Karyotaki, M., Drigas, A., & Skianis, C. (2024). Mobile/VR/Robotics/IoT-Based Chatbots and Intelligent Personal Assistants for Social Inclusion. *International Journal of Interactive Mobile Technologies*, *18*(8). **https://doi.org/10.3991/ijim.v18i08.46473**

Kemper, C. J., Lutz, J., & Margraf-Stiksrud, J. (2012). Eine Ein-Item-Skala zur Einschätzung von Attraktivität: Das Attraktivitätsrating (AR1). *GESIS*.

Kerkhoven, A. H., Russo, P., Land-Zandstra, A. M., Saxena, A., & Rodenburg, F. J. (2016). Gender stereotypes in science education resources: A visual content analysis. *PLoS ONE, 11*(11). **https://doi.org/10.1371/journal.pone.0165037**

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. *Proceedings of the 30th European Conference on Information Systems (ECIS),* 18. **https://aisel.aisnet.org/ecis2022_rp/18**

Kizilcec, R. F., & Kambhampaty, A. (2020). Identifying course characteristics associated with sociodemographic variation in enrollments across 159 online courses from 20 institutions. *PloS One, 15*(10), **https://doi.org/10.1371/journal.pone.0239766**

Kizilcec, R. F., & Saltarelli, A. J. (2019). Psychologically inclusive design: cues impact women's participation in STEM education. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1-10. **https://doi.org/10.1145/3290605.3300704**

Krämer, N. C., Karacora, B., Lucas, G., Dehghani, M., Rüther, G., & Gratch, J. (2016). Closing the gender gap in STEM with friendly male instructors? on the effects of rapport behavior and gender of a virtual agent in an instructional interaction. *Computers and Education, 99*, 1–13. **https://doi.org/10.1016/j.compedu.2016.04.002**

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems, 17*(5), 489–504. **https://doi.org/10.1057/ejis.2008.40**

Kuhail, M. A., Alturki, N., Alramlawi, S., & Alhejori, K. (2022). Interacting with educational chatbots: A systematic review. *Education and Information Technologies, 28*(1), 973–1018. **https://doi.org/10.1007/s10639-022-11177-3**

Kumi–Yeboah, A., Dogbey, J., & Yuan, G. (2017). Exploring factors that promote online learning experiences and academic self-concept of minority high school students. *Journal of Research on Technology in Education, 50*(1), 1–17. **https://doi.org/10.1080/15391523.2017.1365669**

Kuttal, S. K., Sedhain, A., & AuBuchon, J. (2021). Designing a gender-inclusive conversational agent for pair programming: An empirical investigation. *International Conference on Human-Computer Interaction*, 59–75. **https://doi.org/10.1007/978-3-030-77772-2_4**

Latham, A. (2022). Conversational Intelligent Tutoring Systems: The State of the Art. Women in Computational Intelligence: Key Advances and Perspectives on Emerging Topics, 77–101. **https://doi.org/10.1007/978-3-030-79092-9_4**

Leider, A., & Strobel, A. (2020). Using Self-Confidence and Identity to Build Perseverance in MOOC for STEM. *ICERI2020 Proceedings*, 9788–9793. **https://doi.org/10.21125/iceri.2020.2191**

Lobo, I., Rato, D., Prada, R., & Dignum, F. (2022). Socially Aware Interactions: From Dialogue Trees to Natural Language Dialogue Systems. International Workshop on Chatbot Research and Design, *LNCS 13171*, 124–140. **https://doi.org/10.1007/978-3-030-94890-0_8**

Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-Based Digital Assistants: Opportunities, Threats, and Research Perspectives. *BISE, 61*(4), 535–544. **https://doi.org/10.1007/s12599-019-00600-8**

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A survey on evaluation methods for chatbots. *Proceedings of the 2019 7th International conference on information and education technology*, 111–119. **https://doi.org/10.1145/3323771.3323824**

Martha, A. S. D., & Santoso, H. B. (2019). The design and impact of the pedagogical agent: A systematic literature review. *Journal of Educators Online, 16*(1), n1.

McDonnell, M., & Baxter, D. (2019). Chatbots and Gender Stereotyping. *Interacting with Computers,* 31(2), 116–121. **https://doi.org/10.1093/iwc/iwz007**

McKown, C., & Weinstein, R. S. (2003). The development and consequences of stereotype consciousness in middle childhood. *Child development, 74*(2), 498–515.

Meier, P., Beinke, J. H., Fitte, C., Schulte To Brinke, J., & Teuteberg, F. (2021). Generating design knowledge for blockchain-based access control to personal health records. *Information Systems and E-Business Management, 19*, 13–41. **https://doi.org/10.1007/s10257-020-00476-2**

Morana, S., Gnewuch, U., & Jung, D. (2020). The Effect of Anthropomorphism on Investment Decision-Making with Robo-Advisor Chatbots. *Proceedings of the 28th European Conference on Information Systems (ECIS)*. **https://aisel.aisnet.org/ecis2020_rp/63**

Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization, 17*(1), 2–26.**https://doi.org/10.1016/j.infoandorg.2006.11.001**

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**

National Science Foundation. (2019). Women, Minorities, and Persons with Disabilities in Science and Engineering. **https://ncses.nsf.gov/pubs/nsf19304/digest/field-of-degree-women**

Nowak, G. J., Shamp, S., Hollander, B., & Cameron, G. T. (1999). Interactive media: A means for more meaningful advertising. *Advertising and the World Wide Web*, 99–117.

Oliveira, E., de Barba, P. G., & Corrin, L. (2021). Enabling adaptive, personalised and context-aware interaction in a smart learning environment: Piloting the iCollab system. *Australasian Journal of Educational Technology, 37*(2), 1–23. **https://doi.org/10.14742/ajet.6792**

Ozogul, G., Johnson, A. M., Atkinson, R. K., & Reisslein, M. (2013). Investigating the impact of pedagogical agent gender matching and learner choice on learning outcomes and perceptions. *Computers & Education, 67*, 36–50. **https://doi.org/10.1016/j.compedu.2013.02.006**

Rosenberg-Kima, R. B., Plant, E. A., Doerr, C. E., & Baylor, A. L. (2010). The influence of Computer-based model's race and gender on female students' attitudes and beliefs towards engineering. *Journal of Engineering Education, 99*(1), 35–44. **https://doi.org/10.1002/j.2168-9830.2010.tb01040.x**

Schlimbach, R., Stoppel, A., Lampka, L., & Robra-Bissantz, S. (2023). A Picture is Worth a Thousand Words–Exploring Bias in Inclusive Chatbots. *Wirtschaftsinformatik 2023 Proceedings,* 12.

Scholten, M. R., Kelders, S. M., & van Gemert-Pijnen, J. E. W. C. (2017). A scoped review of the potential for supportive virtual coaches as adjuncts to self-guided web-based interventions. *Persuasive Technology: Development and Implementation of Personalized Technologies to Change Attitudes and Behaviors: 12th International Conference*, *LNCS 10171*, 43–54. **https://doi.org/10.1007/978-3-319-55134-0_4**

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797–811. **https://doi.org/10.1037/0022-3514.69.5.797**

Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., & Churchill, E. (2020). Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends® in Human–Computer Interaction, 13*(1), 1–69. **http://dx.doi.org/10.1561/1100000056**

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European Journal of Information Systems, 25*, 77–89.

Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the shoulders of giants: Challenges and recommendations of literature search in information systems research. *Communications of the Association for Information Systems, 37*(1), 9. **https://doi.org/10.17705/1CAIS.03709**

Wellnhammer, N., Dolata, M., Steigler, S., & Schwabe, G. (2020). Studying with the help of digital tutors: Design aspects of conversational agents that influence the learning process. Proceedings of the 53rd Hawaii International Conference on System Sciences.

Yildiz, E., Bensch, S., & Dignum, F. (2021). Incorporating Social Practices in Dialogue Systems. *International Workshop on Chatbot Research and Design*, *LNCS 13171*, 108–123. **https://doi.org/10.1007/978-3-030-94890-0_7**

Yücel, Y., & Rızvanoğlu, K. (2019). Battling gender stereotypes: A user study of a code-learning game,"Code Combat," with middle school children. *Computers in Human Behavior, 99*, 352–365.

Zabel, S., & Otto, S. (2021). Bias in, bias out–the similarity-attraction effect between chatbot designers and users. *Human-Computer Interaction International 2021*, 184–197.

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, *2*, **https://doi.org/10.1016/j.caeai.2021.100025**