The influence of avatar interfaces on argumentative dialogues

Annalena Aicher, Klaus Weber, Elisabeth André, Wolfgang Minker, Stefan Ultes

Angaben zur Veröffentlichung / Publication details:

Aicher, Annalena, Klaus Weber, Elisabeth André, Wolfgang Minker, and Stefan Ultes. 2023. "The influence of avatar interfaces on argumentative dialogues." In IVA '23: proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents, Würzburg, Germany, September 19-22, 2023, edited by Birgit Lugrin, Marc Latoschik, Sebastion von Mammen, Stefan Kopp, Florian Pécune, and Catherine Pelachaud, 24. New York, NY: Association for Computing Machinery (ACM). https://doi.org/10.1145/3570945.3607343.

The Influence of Avatar Interfaces on Argumentative Dialogues

Annalena Aicher
annalena.aicher@uni-ulm.de
Institute of Communications Engineering, Ulm University
Ulm, Germany
Ubiquitous Computing Systems Laboratory, NAIST
Ikoma, Nara, Japan

Klaus Weber
Elisabeth André
klaus.weber@uni-a.de
elisabeth.andre@uni-a.de
Human-Centered Artificial Intelligence, University of Augsburg
Augsburg, Germany

Wolfgang Minker
wolfgang.minker@uni-ulm.de
Institute of Communications Engineering, Ulm University
Ulm, Germany

Stefan Ultes
stefan.ultes@uni-bamberg.de
Natural Language Generation and Dialogue Systems, University of Bamberg
Bamberg, Germany

ABSTRACT

Humans form opinions and justify different points of view by exchanging arguments and knowledge. Likewise to human-human interaction, the way arguments are presented influence the user's willingness to engage into a critical reflection. Especially when interacting with conversational agents the user's engagement and motivation are important factors and highly influence the success or failure of such a mixed team. To maintain the users' trust and satisfaction, the users' perception of the respective system is an important indicator. Thus, this work investigates the design of a cooperative argumentative dialogue system using a virtual avatar compared to a non-avatar interface by evaluating a crowdsourcing study conducted with 84 participants. The results indicate, that the avatar system is perceived as significantly more appealing and natural and thus, engaging which also influences the acceptance and perception of the quality of presented arguments. Furthermore, we found that the presence of the avatar often led to an increase in the anticipated level of conversational proficiency similar to that of a human interlocutor. Therefore, this work provides important insights for the design of future cooperative argumentative virtual avatar interfaces.

CCS CONCEPTS

• **Human-centered computing** → Empirical studies in interaction design; User interface design; *Empirical studies in HCI*; *Graphical user interfaces*; Text input; **User studies**.

KEYWORDS

Conversational Engagement, User Trust, Avatar Interface, Human-Computer Interaction, Crowdsourcing Study

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607343

ACM Reference Format:
Annalena Aicher, Klaus Weber, Elisabeth André, Wolfgang Minker, and Stefan Ultes. 2023. The Influence of Avatar Interfaces on Argumentative Dialogues. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3570945.3607343

1 INTRODUCTION

In recent years the popularity of virtual assistants or agents or avatars has increased rapidly. This development is fueled by advancements in computer/digital technologies as well as the growing importance of online service experiences, such as education, gaming, banking, and shopping [11, 14]. For example, a variety of companies, including Amazon.com, Booking.com, and Yahoo!, already offer their online customers virtual salesperson or recommendation avatar[33] services to assist in online shopping [17]. Miao et al. [19] define avatars as "digital entities with anthropomorphic appearance, controlled by a human or software, with an ability to interact". Please note, that within this paper we use the term avatar as the human-like 3D representation of a conversational agent which personifies our argumentative dialogue system.

Recent findings in marketing [19] or in games research have shown that the use of player avatars is effective in improving interest, enjoyment, and other intrinsic motivation aspects [27]. Qiu et al. [27] claim conversational agents have advantages over traditional graphical user interfaces as they allow for a more human-like interaction. For example, Rebolledo-Mendez et al. [30] show that the use of avatars in Computer-Aided Instruction has an intrinsically motivating effect on learners. Even though the use of avatars displays a great potential, their effectiveness varies significantly [19]. Lin et al. [17] point out that discrepancies between online customer reviews and the purchase recommendations offered by the virtual salesperson affects customers' trust and their willingness to follow the avatar's recommendations.

Furthermore, as the literature in this domain is very fragmented [19] it is unclear whether avatars have an impact on the course of argumentative debates [7]. This raises the question of whether the usage of a virtual avatar as a counterpart in an argumentative discussions is perceived as motivating and engaging. This is especially important as we aim for the user to scrutinize arguments thoroughly to get a

well-founded opinion. Therefore, this paper aims to shed light on the influence of avatars in cooperative argumentative dialogues and describes the impact on the user's engagement, trust and willingness to scrutinize arguments (critically).

The remainder of the paper is as follows: after an overview of related work in Section 2, Section 3 describes the architecture of the argumentative dialogue system (ADS), especially with respect to the graphical user interface (GUI). Section 4 covers the experiment and study setup. Afterwards, Section 5 describes the evaluation results. The discussion of the results is included in Section 6 followed by a brief conclusion and outlook on future work in Section 7.

2 RELATED WORK

In state-of-the-art literature avatars in virtual environments [10, 25, 26, 34], their control [4, 38], and their personalization [9] have been extensively studied.

Whereas we focus on the influence of a virtual avatar as a counterpart, a lot of research in this field aims to analyse the influence of self-identification with a virtual avatar representing the user in a virtual space. For instance, the findings of Ratan et al. [29] support the claim that the integration of avatars in learning contexts enhances the student motivation and performance. They define the term "avatarification"as the "utilization of virtual self-representations within a mediated environment in order to facilitate interactions in that environment". This is underpinned by several findings Ducheneaut et al. [9], Mohd Tuah et al. [22] which show that avatar personalization positively affects players' emotions and thus, enhances the individual's motivation and game experience. Similar positive effect were described by Qiu et al. [27] pointing out the potential of worker avatars to improve worker satisfaction in microtask crowdsourcing.

In general, many studies in research field of computer-mediated communication have shown that higher aesthetic and behavioral realism leverages user engagement, leads to higher user acceptance and sense of "social copresence" Bickmore and Mauer [6], Kang et al. [12, 13], Wang et al. [36]. Apart from the choice and design of the avatar, i.e. its gender [16, 37, 39], the results of Aseeri and Interrante [3] imply that the visual and nonverbal cues afforded by different avatar representations (no avatar, scanned avatar, real avatar) in the context of several cooperative tasks affect the user experience. A positive effect on the communication experience is also described by Rincon-Nigro [32] even if synthetic talking avatars do not appear completely realistic. This is strengthened by Miyake and Ito [21] stating that a virtual conversational agent in a spoken dialogue system increases the response accuracy, the feeling of vividness of conversation and easiness to talk to the system.

Still the literature focusing on the influence of avatars in argumentative dialog systems is very scarce. Blount et al. [7] present an approach on how participants can shape or craft their own avatar appearance to suit their needs/purpose and impact the course of an argumentative debate in the virtual sphere. Furthermore, Mikeska et al. [20] examine the perceptions and engagement of secondary preservice teachers within an online, simulated classroom with middle school student avatars to train and practice argumentation-focused scientific discussions. To the best of our knowledge aforementioned findings still lack an analysis of the change in engagement, motivation and perception of a cooperative argumentative dialogue system when a virtual human-like avatar instead of a chat-based interface is employed. Within this paper we aim to close this gap and furthermore analyse whether a virtual conversational agent influences the trustworthiness [31] and the respective implications for a more thorough argument reflection.

3 ADS ARCHITECTURE

In the following, the architecture of our ADS and its components, in particular the underlying dialogue model, argument structure and interface are outlined.

3.1 Dialogue Model and Argument Structure

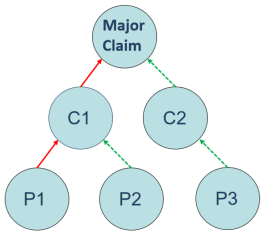

Figure 1: Visualization of argument tree structure. The discussion topic is the root node, which is supported by the claim C2 (green dotted arrow) and attacked by claim C1 (red solid arrow). The respective leaf nodes are premises P1, P2 and P3.

}

To be able to combine our ADS with existing argument mining approaches to ensure its flexibility in view of discussed topics, we follow the bipolar argument annotation scheme introduced by Stab and Gurevych [35][1]. It distinguishes three different types of components (Major Claim, Claim, Premise), which are structured in the form of bipolar argumentation trees depicted in Figure 1. The overall topic of the debate is formulated as the *Major Claim* representing the root node in the graph. *Claims* (C1 and C2) on the other hand are assertions which formulate a certain opinion targeting the *Major Claim* but still need to be justified by further arguments, *premises* (P1 and P2) respectively. We consider two relations between these argument components (nodes): *support* (green dotted arrows) or *attack* (red solid arrows). Each component apart from the Major Claim (which has no relation) has exactly one unique relation to another component. This leads to a non-cyclic tree structure, where each node or "parent" (C1 and C2) is supported or attacked by its "children". If no children exist, the node is a leaf (e.g. $P1, P2$ and $P3$) and marks the end of a branch. The interaction between the system and the user is separated in turns, consisting of a user action and corresponding natural language answer of the system. The system response is based on the original textual representation of the argument components, which is embedded in moderating utterances. Table 1 shows the possible moves (actions) the user can perform. These enable the user to navigate through the argument tree and enquire more information. Furthermore, the users are able to give

[1]Due to the generality of the annotation scheme, the system is not restricted to the herein considered data. In general, every argument structure that can be mapped onto the applied scheme can be used.

feedback, whether they agree or disagree with the given argument by clicking the respective button on the interface (see Section 3.2). To exit the conversation, the user presses the "Finish" button (after listening to the minimum of required arguments). In this ADS a sample debate on the topic *Marriage is an outdated institution* provides a suiting argument structure[2]. It serves as knowledge base for the arguments and is taken from the *Debatabase* of the idebate.org[3] website. It consists of a total of 72 argument components (1 *major claim*, 10 *claims* and 61 *premises*) and their corresponding relations and is encoded in an OWL ontology [5] for further use. In each "why pro/con" move a single argument component is presented to the user. To prevent the user from being overwhelmed by the amount of information, the available arguments are presented to the users incrementally on their request.

Table 1: Description of possible user actions.

Move	Description
why_{pro}	Request for a pro argument.
why_{con}	Request for a con argument.
$level_{up}$	Returns to the parent node.
$help$	Request for help what to do next.

3.2 Interface

The non-avatar system's GUI is illustrated in Figure 2 and the avatar system's GUI in Figure 3. The interfaces are based and extended from two argumentative dialogue systems (ADS) introduced by Aicher et al. [2], Rach et al. [28][4]. Apart from the different GUI with different output modalities (chat-based output or a spoken avatar output) both systems are completely identical, especially with regard to the systems dialogue strategy and response generation. The avatar interface is based on the Charamel™ avatar [5] which presents the system utterance via synthetic speech by utilizing Nuance TTS and Amazon Polly Voices[6]. We opted for a full-body representation of the avatar as it moves across the screen to introduce and highlight various elements of the GUI. Furthermore, our focus in the study was to examine the fundamental influence of an avatar using speech gestures, which emphasized the verbal expressions and presentation of arguments and did not include the analysis of the influences of explicit facial expressions/emotions. Instead of the avatar, in the non-avatar interface the dialogue history is shown in the center of the screen without speech output. The visualization of the dialogue history shows the system's responses left-aligned in green and corresponding user moves right-aligned in blue (see Figure 2). In the avatar interface the dialogue history is shown on the right side of the screen. Furthermore, on the left side, the sub-graph of the bipolar argument tree structure (with the displayed claim as root) is shown. The current position (i.e., argument) is displayed with a white node outlined with a blue line. Already heard arguments are shown in blue. Nodes shown in grey are still unheard. A progress bar at the top of the screen shows the number of arguments that were already discussed and how many are still unknown to the user at each stage of the interaction. Additionally, both interfaces include buttons for the users to state their agreement or disagreement regarding the currently presented argument (see Figure 2 and 3). Furthermore the 'Finish' button can be clicked to end the interaction if the minimum number of arguments (10) have been heard. Below these buttons a chat-input line is shown where the users are able to type in their answer.

The user's typed input is processed by an NLU framework [1]. Its intent classifier uses the BERT Transformer Encoder presented by Devlin et al. [8] and a bidirectional LSTM classifier. The system-specific intents are trained with a set of sample utterances of previous user studies. The response generation is based on the original textual representation of the argument components. The annotated sentences are slightly modified to form a stand-alone utterance serving as a template for the respective system response. Additionally, a list of natural language representations for each system move was defined. During the generation of the utterances, the explicit formulation and introductory phrase are randomly chosen from this list.

4 USER STUDY SETTING

We conducted a user study with 84 participants (aged 18-65; 52 female, 31 male, 1 "other/do not want to say") divided into two groups: 46 participants interacted with a virtual avatar interface further on called "avatar group", and 38 participants with a non-avatar interface further on called "non-avatar group". Our descriptive analysis aims to analyse the influence of virtual avatar on the user opinion, engagement and trust. Furthermore, we describe the impact of an avatar interface on the user's perception of an argumentative dialogue system. The study was conducted online via the crowdsourcing platform "Crowdee" (https://www.crowdee.com/, 31st May -23rd June 2022) with participants from the UK, US, and Australia (English native speakers to avoid language barrier effects) without a topic-specific background. All participants were given an introductory text explaining the underlying argument structure and, especially, how to interact with the ADS including a test question the participants had to answer correctly to proceed with the study. The users were asked to explore enough arguments to build a well-founded opinion on the topic *Marriage is an outdated institution*. The participants were not told anything about the purpose to investigate their conversational engagement but only to select at least ten arguments.

Before the conversation some demographic data was collected, as well as the user's opinion and interest (5-point Likert scale) in the topic. After the conversation the participants had to rate statements on a 5-point Likert scale (1 = totally disagree, 5 = totally agree) concerning the interaction. They were taken from a questionnaire according to ITU-T Recommendation P.851[8] [24]. Furthermore, we asked the users about their trust towards the system using the

[2]We considered this topic as suitable as topics with a "more substantial societal need" are much more likely to cause strong emotions and biases due to their relevance and timeliness. We aimed to minimize these effects to better differentiate between the influences attributed to the topic itself and those associated with the avatar interface.

[3]https://idebate.org/debatabase (last accessed 23th July 2021). Material reproduced from www.iedebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

[4] Aicher et al. [2] evaluated the likeability and motivation of users to interact via speech or drop-down menu with an argumentative dialogue system to explore a controversial topic. Rach et al. [28] introduced an argumentative dialogue system in which two virtual agents discuss a specific topic in order to convince the user of their stance.

[5]https://www.charamel.com/competence/avatare

[6]https://docs.aws.amazon.com/polly/latest/dg/voicelist.html

[8]Such questionnaires can be used to evaluate the quality of speech-based services.

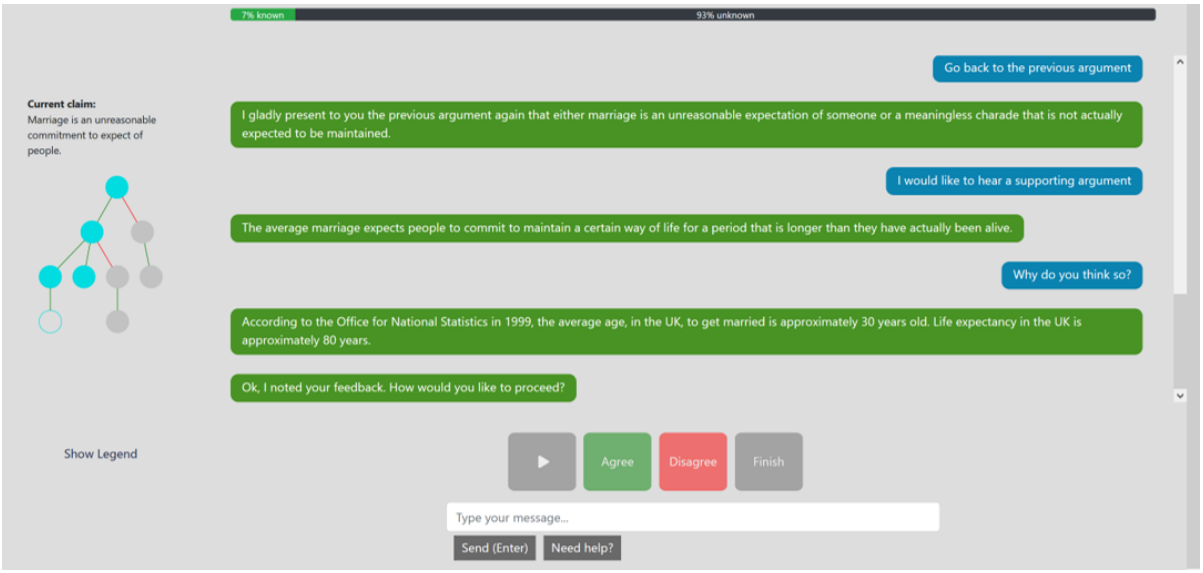

Figure 2: GUI without avatar. Above the chat-input line, four buttons are shown - a play button to start the interaction, one 'Agreement' and one 'Disagreement' button to state one's opinion towards the current argument and a 'Finish' button to end the conversation. Above those buttons the dialogue history is shown. The system utterances are marked in green, user responses in blue. On the left side, the sub-graph of the current branch is visible.

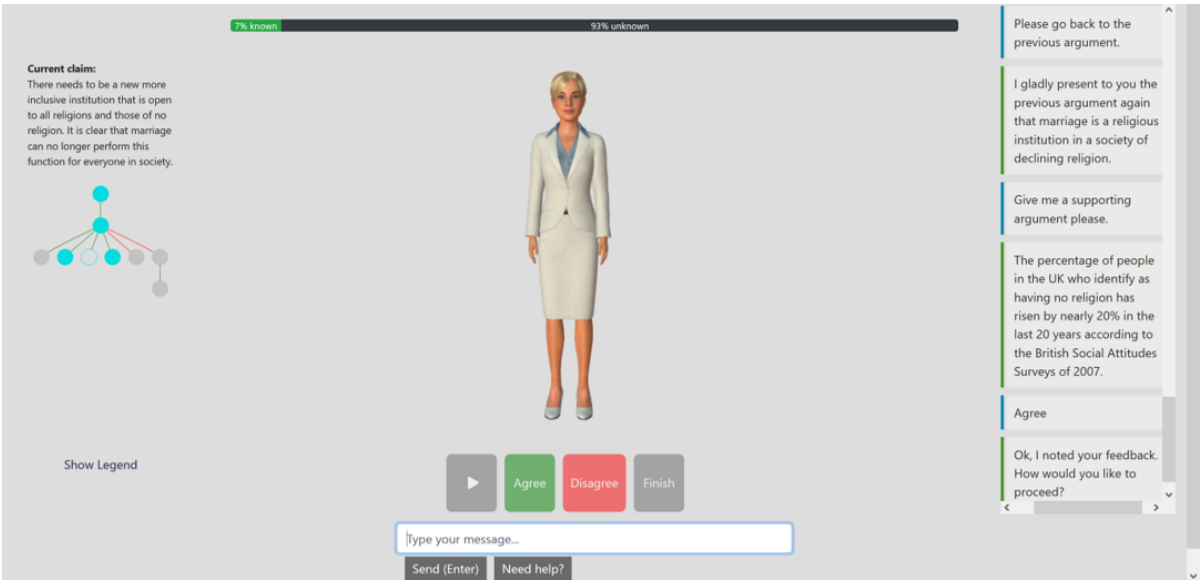

Figure 3: GUI with avatar. Above the chat-input line four buttons are shown similar to the non-avatar GUI. In contrast to the latter, above those buttons the virtual avatar is shown and the dialogue history is placed on the right side of the screen. The system utterances are marked in green and user responses in blue. On the left side, the sub-graph of the current branch is visible.

questionnaire of Körber [15][9] consisting of six items and 12 items introduced by the questionnaire of O'Brien et al. [23] to investigate the conversational engagement.

During the study, we collected the following data anonymously[10]: general statistics, Dialogue history (system and user utterances), user's agreement/disagreement to arguments (indicated by user clicking on the respective button).

[9]This questionnaire was developed of to measure trust in automation.

[10]In accordance with the applicable official privacy policy to which each user has voluntarily agreed prior to the study.

Table 2: Means and standard deviations of the questionnaire items regarding the user's perception of the system P.851 [24] grouped by the following aspects: information provided by the system (IPS), communication with the system (COM), system behaviour (SB), dialogue (DI), user's impression of the system (UIS), acceptability (ACC), and argumentation (ARG). Significant differences are indicated by a bold p value.

		Avatar		Non-Avatar		
Asp.	**Question**	*M*	*SD*	*M*	*SD*	*p* value
IPS	1. The system has provided you the desired information.	**3.44**	0.90	3.36	1.01	0.781
	2. The system's answers and proposed solutions were clear.	3.54	1.01	**3.67**	0.93	0.585
	3. You would rate the provided information as true.	**3.67**	0.75	3.61	0.84	0.749
	4. The information provided by the system was complete.	3.25	1.00	**3.36**	0.93	0.794
COM	1. The system always understood you well.	2.67	1.098	**2.75**	1.296	0.848
	2. You had to concentrate in order to understand what the system expected from you.*[7]	3.71	0.922	**3.83**	0.971	0.501
	3. The system's responses were well understandable.	**3.75**	0.937	3.54	0.944	0.316
	4. You were able to interact efficiently with the system.	2.92	1.048	**3.08**	1.204	0.533
SB	1. You knew at each point of the interaction what the system expected from you.	**2.71**	1.021	2.69	1.009	0.774
	2. In your opinion, the system processed your responses (specifications) correctly.	3.23	0.905	**3.28**	1.111	0.665
	3. The system's behavior was always as expected.	**3.02**	0.934	2.89	1.090	0.410
	4. The system often failed to understand you.*	**3.27**	1.233	3.00	1.195	0.341
	5. The system reacted naturally.	**3.27**	1.067	2.78	0.959	**0.037**
	6. The system reacted flexibly.	**3.02**	1.101	2.75	1.131	0.255
	7. You were able to control the interaction in the desired way.	2.81	1.003	**2.92**	1.156	0.580
	8. The system reacted too slowly.*	3.29	1.091	**2.19**	0.980	**<0.001**
	9. The system reacted politely.	**4.37**	0.606	4.31	0.624	0.615
	10. The system's responses were too long.*	2.58	1.069	**2.03**	0.696	**0.016**
DI	1. You perceived the dialogue as natural.	**3.52**	0.899	3.03	1.028	**0.032**
	2. It was easy to follow the flow of the dialogue.	3.31	0.926	**3.36**	1.046	0.869
	3. The dialogue was too long.*	2.42	0.895	**2.11**	0.747	0.093
	4. The course of the dialogue was smooth.	3.44	0.796	3.44	0.939	0.864
	5. You and the system could clear misunderstandings easily.	2.79	1.010	**3.00**	1.146	0.445
	6. You would have expected more help from the system.*	3.75	1.042	**3.19**	1.064	**0.014**
UIS	1. Overall, you were satisfied with the dialogue.	**3.40**	1.067	3.22	1.124	0.431
	2. The dialogue with the system was useful.	3.27	1.1162	**3.39**	1.022	0.686
	3. It was easy for you to obtain the information you wanted.	**2.92**	1.007	2.86	1.046	0.804
	4. You have perceived the dialogue as pleasant.	**3.90**	0.805	3.56	1.046	0.154
	5. You felt relaxed during the dialogue.	3.42	1.007	**3.69**	1.064	0.146
	6. Using the system was fun.	**3.19**	1.142	2.83	1.207	0.203
ACC	1. In the future, you would use the system again.	3.83	0.81	**3.86**	0.723	0.892
	2. You would would recommend the system to a friend.	**3.21**	1.184	2.75	1.105	0.067
ARG	1. I felt motivated by the system to discuss the topic.	**3.64**	0.988	2.94	1.241	0.068
	2. I would rather use this system than read the arguments in an article.	**3.15**	1.288	2.94	1.372	0.504
	3. The possible options to respond to the system were sufficient.	3.00	1.111	**3.06**	1.264	0.806
	4. The arguments the system presented are conclusive.	**3.21**	1.031	3.06	0.924	0.419
	5. I felt engaged in the conversation with the system.	**3.40**	1.267	2.83	1.159	**0.039**
	6. The interaction with the system was confusing.*	**2.73**	1.198	3.14	1.376	0.149
	7. I do not like that the arguments are provided incrementally.*	3.04	1.267	**2.67**	1.014	0.111

5 RESULTS

In the following, we describe the results of the previously described user study. On average the participants interacted significantly longer ($p = 0.013$, $r = 0.34$) with the ADS for 32:31 minutes (SD: 19:31) in the avatar and for 20:57 (SD: 9:34) minutes in the non-avatar setting. For the evaluation of the self-assessment questionnaire the mean M and standard deviation SD were determined for each single item and system. Regarding all items the assumption of a normal distribution based on the Shapiro-Wilk-Test had to be discarded ($W = 0.696 - 0.913$, $p < 0.001$). Thus, to determine whether the

difference between the two systems means Δ_M is significant, we used the non-parametric Mann-Whitney U test [18] for two independent samples with no specific distribution. Please note, that We conducted an explorative study and thus refrained from employing multiple test correction methods. The category "Overall Quality" ("What is your overall impression of the system?") is not included in Table 2 as it is rated on a different 5-point Likert scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad). Our analysis shows a statistically significant ($p = 0.047$) difference between the two systems. The avatar system received an averaged rating of 3.67 (SD 0.930), outperformed the non-avatar system with 3.17 (SD 1.082). This difference is considered moderate, as indicated by the effect size $r = 0.216$.

As shown in Table 2 the single item analysis between both groups does not show any significant differences regarding the aspects "information provided by the system" (IPS), "communication with the system" (COM), "user's impression of the system" (UIS) and "acceptability" (ACC).

Table 3: Means and standard deviations of the questionnaire items regarding user trust Körber [15] grouped by the following aspects: understanding/predictability (UP), familiarity (F), propensity to trust (PT) and trust in automation (TA). Significant differences are indicated by a bold p value.

		Avatar		Non-Avatar		
Asp.	**Question**	*M*	*SD*	*M*	*SD*	p value
UP	The system state was always clear to me.	**2.90**	1.036	2.67	0.894	0.244
	The system reacts unpredictably.*	**2.50**	0.875	2.44	1.107	0.455
	I was able to understand why things happened.	3.10	0.994	**3.25**	0.906	0.522
	It's difficult to identify what the system will do next.*	3.33	1.018	**3.06**	1.145	0.281
F	I already know similar systems.	**2.58**	1.069	2.50	1.082	0.704
	I have already used similar systems.	2.50	1.072	**2.56**	1.107	0.821
PT	One should be careful with unfamiliar automated systems.*	**3.46**	0.967	3.97	0.810	**0.025**
	I rather trust a system than I mistrust it.	3.19	0.891	3.19	0.822	0.969
	Automated systems generally work well.	**3.25**	1.000	2.72	1.085	**0.022**
TA	I trust the system.	**3.44**	0.823	2.97	1.028	**0.039**
	I can rely on the system	**2.96**	0.898	2.94	1.040	0.951

A significant difference is notable regarding the aspect "system behaviour" (SB) in two single items (natural system reaction (SB 5[11]), system's response speed (SB 7) and length (SB 10)). With regard to the aspect "Dialogue" two single items (naturalness of the dialogue (DI 1) and expected help (DI 6)) showed a significant difference between the two groups (effect sizes: $r_{\text{DI 1}} = 0.234$, $r_{\text{DI 6}} = 0.269$). With regard to our self-added aspect "argumentation" (ARG), we observed a significant difference ($r_{\text{ARG 5}} = 0.226$) in the single item "engagement induced by the system".

When the single items or their inverted counterparts marked with (*) are summarized in their associated aspects, there is no significant difference (p=0.290-0.993) perceivable for any of these merged aspects.

[11]The numbers following the abbreviations correspond to the respective question numbers.

Table 4: Means and standard deviations of the items of the short user engagement questionnaire O'Brien et al. [23] grouped by the following aspects: Focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and RW. Significant differences are indicated by a bold p value.

		Avatar		Non-Avatar		
Asp.	**Question**	*M*	*SD*	*M*	*SD*	p value
FA	I lost myself in this experience.	**2.63**	1.160	2.19	0.856	0.121
	The time I spent using the application just slipped away.	**2.60**	1.047	2.44	1.054	0.451
	I was absorbed in this experience.	**3.08**	1.182	2.69	1.117	0.144
PU	I felt frustrated while using the application.*	**3.10**	1.242	3.22	1.245	0.641
	I found this application confusing to use.*	**2.98**	1.176	3.19	1.305	0.385
	Using this application was taxing.*	**2.71**	1.166	2.97	1.253	0.379
AE	The application was attractive.	**3.19**	0.842	3.14	0.833	0.875
	The application was aesthetically appealing.	**3.25**	0.838	3.06	0.791	0.174
	This application appealed to my senses.	**2.98**	0.863	2.89	0.887	0.615
RW	Using the application was worthwhile.	**3.38**	1.003	2.86	1.046	**0.022**
	My experience was rewarding.	**3.40**	1.162	2.94	1.094	0.076
	I felt interested in this experience.	**3.71**	1.071	3.56	1.157	0.586

In Table 3 the single item results for the questionnaire [15] are shown to analyse the user trust regarding the ADS. In three single items a significant difference with a moderate effect size ($r_{\text{PT 3}} = 0.244$, $r_{\text{PT 1}} = 0.250$, $r_{\text{TA 1}} = 0.244$) between the avatar and the non-avatar group is perceived. Merging the single items (inverted counterparts respectively) into their associated four aspects (UP, F, PT, TA), results in a significant difference for PT ($p = 0.015$, $r = 0.266$).

Table 4 displays the results of the short form of the user engagement scale introduced by O'Brien et al. [23]. Even though only one of the single items did show a significant difference with a moderate effect size ($r_{\text{RW 1}} = 0.250$), the avatar system was rated better than the non-avatar one in every item. When merging the single items (inverted counterparts respectively) into their associated four aspects (FA, PU, AE, RW) no significant difference ($p = 0.061 - 0.358$) could be perceived.

6 DISCUSSION

In the following section we discuss the results of our study presented in Section 5. In general, participants interacted significantly longer with the avatar. This can be explained by the fact that listening to the spoken utterance of the avatar, and re-reading the respective response in the dialogue history takes longer than just reading the displayed answer. Furthermore, sometimes the reaction time of the avatar system was longer compared to the non-avatar system due to lags in the connection with the avatar server.

Even though the differences between the avatar and non-avatar group regarding the merged aspects IPS, COM, SB, DI, UIS, ACC

and ARG (see Table 2) are insignificant, we can perceive some consistent, aspect-overlapping tendency. Especially regarding the perceived naturalness (SB 4, DI 1) a significant influence of the avatar system compared to the non-avatar system is observed. These findings imply that an avatar interface seems to influence the user's perception of naturalness in course of an argumentative dialogue, which underpins the claim in state-of-the-art literature, that avatars can be used to design a more human-like, natural conversation. Furthermore, we observe that the users felt more engaged in the avatar system (ARG 5). In detail this is explored with the items shown in Table 4. It clearly can be seen that the avatar system seems to have an engaging effect on the user, which is significant for the impression that using the system was worthwhile (RW 1). Still the rating of perceived usability (PU) indicate there is need for enhancement, especially regarding the errors of the ASR and the explanation of the system's reaction, if the user is not understood correctly. This is also observable in Table 2 where system's reaction time (SB 8) and responses (SB 10) were rated significantly worse in the avatar system. Unfortunately, this delay is caused by the avatar itself, because an external server access is necessary. Depending on the connection quality and load, this results in different system response times. In contrast, the response of the non-avatar setting is only generated internally by our ADS and presented immediately on the GUI.

The significant difference in the expected help the system should have provided (DI 6) implies that the avatar on one hand side tends to raise the expectation to the one of a human conversational partner and on the other if these expectations are met could lead to a significantly stronger acceptance comparable to a human conversational partner. This is underpinned by the voluntary free-text remarks of participants on the study and the system. Participants in the non-avatar setting mainly commented on the performance, technical limitations of the ADS and suggested potential ways to enhance it (e.g. "the system might as well have been a list of pros and cons...", "speed up the results would make the experience better" or "it would be good if there are more buttons representing the options to interact with BEA, or a drop-down menu of the previously typed words."). In the avatar setting participants focused their comments more on the overall impression of the system and suggestions for improvement that increase the naturalness and flexibility of the dialogue ("I think the system needs more detailed arguments not just stating statistics" or "[...] The system moved very smoothly and was engaging. I found this study extremely interesting and very valid as we move toward more automated systems."). The results in Tab. 3 implicate that users seem to have a slight tendency to trust the avatar system more than the non-avatar one, especially regarding the propensity to trust. The fact that avatar participants are not significantly more familiar with similar systems, implies that an avatar interface in ADS can help the user to intuitively access arguments and influences the user trust positively. This finding supports our suggestion to use an avatar interface in ADS and could certainly help to increase user trust by individualizing the avatar.

This study, however, is subject to three limitations that could be addressed in future research. First, we do not compare different avatar settings personalized to the user, such as gender, age, variation of motion or mimics, which have a great impact on the user's social perception and cognition Wessler et al. [37], Zanbaka et al. [39]. For this explorative study we chose an easily implementable, commonly accessible, representative avatar, rather than the best-fitting one and compared it to a purely chat-based interface to assess the influence of avatars for this argumentative interaction. However, further work in this area should explore the impact of additional features of avatars and the environment in which they are presented [7]. Second, as prior research studies relevant to our observations are very limited and no comparable results exists yet, our quantitative analysis focuses on the participants' self-assessment answers and argument exploration behavior. In future work, the validity of our findings would be strengthened if they could be compared to a baseline and additionally be supported by qualitative analysis (e.g. by free text responses of users). Third, as already mentioned, during the utilization of the Charamel avatar, instances occurred where the time required for a response was perceived by the participants to be excessively delayed. It is crucial that this issue is addressed in subsequent research. Nevertheless, it is important to note that the avatar system exhibited a significantly higher overall quality when compared to its non-avatar counterpart. We believe that this impression of this system could be further enhanced by addressing the aforementioned limitations and adapt the avatar to the needs and preferences of the user.

7 CONCLUSION AND FUTURE DIRECTIONS

In this work we investigated the impact of avatar versus non-avatar interfaces on user perception, engagement, and trust in argumentative dialogue systems. A crowdsourcing study with 84 participants was conducted, and self-assessment questionnaires were used for analysis. Results indicate that the use of the avatar positively influenced user engagement, trust and perception of naturalness in interaction, with a human-like design raising expectations of communication and help, akin to a human conversational partner. However, errors or delays in response time had a more negative impact in the avatar setting.

Given the increasing prevalence of social web, it is crucial to understand the effects of avatars on interpersonal communication, particularly in argumentation [7]. Future work will explore the potential of personalized avatar implementation to enhance interaction in argumentative dialogue systems, through social presence influence, positive feedback, and emotional connection, leading to increased user engagement and satisfaction. Therefore, this work provides important implications for the design of future cooperative argumentative virtual avatar interfaces.

ACKNOWLEDGMENTS

This work has been funded by the DFG within the project "BEA - Building Engaging Argumentation", Grant no. 313723125, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

REFERENCES

[1] Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems* 242 (2022), 108318. https://www.sciencedirect.com/science/article/pii/S0950705122001149

[2] Annalena Aicher, Nadine Gerstenlauer, Isabel Feustel, Wolfgang Minker, and Stefan Ultes. 2022. Towards building a spoken dialogue system for argument

exploration. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 1234–1241.

[3] Sahar Aseeri and Victoria Interrante. 2021. The Influence of Avatar Representation on Interpersonal Communication in Virtual Social Environments. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2608–2617. https://doi.org/10.1109/TVCG.2021.3067783

[4] Francesca Barrientos and John Canny. 2002. Cursive: A novel interaction technique for controlling expressive avatar gesture. (04 2002). https://doi.org/10.1145/502348.502370

[5] Sean Bechhofer. 2009. OWL: Web ontology language. In *Encyclopedia of Database Systems*. Springer, 2008–2009.

[6] Timothy Bickmore and Dan Mauer. 2006. Modalities for building relationships with handheld computer agents. *Conference on Human Factors in Computing Systems - Proceedings*, 544–549. https://doi.org/10.1145/1125451.1125567

[7] Tom Blount, David E. Millard, and Mark J. Weal. 2015. On the Role of Avatars in Argumentation. In *Proceedings of the 2015 Workshop on Narrative & Hypertext* (Guzelyurt, Northern Cyprus) *(NHT '15)*. Association for Computing Machinery, New York, NY, USA, 17–19. https://doi.org/10.1145/2804565.2804569

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Nicolas Ducheneaut, Don Ming-Hui Wen, Nicholas Yee, and Greg Wadley. 2009. Body and mind: A study of avatar personalization in three virtual worlds. *Proceedings of CHI 2009*, 1151–1160. https://doi.org/10.1145/1518701.1518877

[10] Thomas Erickson, N. Sadat Shami, Wendy A. Kellogg, and David W. Levine. 2011. Synchronous Interaction among Hundreds: An Evaluation of a Conference in an Avatar-Based Virtual Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 503–512. https://doi.org/10.1145/1978942.1979013

[11] Marion Garnier and Ingrid Poncin. 2013. The avatar in marketing: Synthesis, integrative framework and perspectives. *Recherche et Applications en Marketing (English Edition)* 28, 1 (2013), 85–115.

[12] Sin-Hwa Kang, H.W.J. Kang, and Sasi Ala. 2008. Communicators' Perceptions of Social Presence as a Function of Avatar Realism in Small Display Mobile Communication Devices. 147 – 147. https://doi.org/10.1109/HICSS.2008.95

[13] Sin-Hwa Kang, James Watt, and Sasi Ala. 2008. Social copresence in anonymous social interactions using a mobile video telephone. *Conference on Human Factors in Computing Systems - Proceedings*, 1535–1544. https://doi.org/10.1145/1357054.1357295

[14] Sara Kim, Rocky Peng Chen, and Ke Zhang. 2016. Anthropomorphized Helpers Undermine Autonomy and Enjoyment in Computer Games. *Journal of Consumer Research* 43, 2 (04 2016), 282–302. https://doi.org/10.1093/jcr/ucw016 arXiv:https://academic.oup.com/jcr/article-pdf/43/2/282/7919248/ucw016.pdf

[15] Moritz Körber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20*. Springer, 13–30.

[16] Mika Lehdonvirta, Yosuke Nagashima, Vili Lehdonvirta, and Akira Baba. 2012. The Stoic Male: How Avatar Gender Affects Help-Seeking Behavior in an Online Game. *Games and Culture* 7, 1 (2012), 29–47. https://doi.org/10.1177/1555412012440307 arXiv:https://doi.org/10.1177/1555412012440307

[17] Yu-Ting Lin, Her-Sen Doong, and Andreas B Eisingerich. 2021. Avatar design of virtual salespeople: Mitigation of recommendation conflicts. *Journal of service research* 24, 1 (2021), 141–159.

[18] Patrick E. McKnight and Julius Najab. 2010. *Mann-Whitney U Test*. American Cancer Society, 1–1.

[19] Fred Miao, Irina V. Kozlenkova, Haizhong Wang, Tao Xie, and Robert W. Palmatier. 2022. An Emerging Theory of Avatar Marketing. *Journal of Marketing* 86, 1 (2022), 67–90. https://doi.org/10.1177/0022242921996646 arXiv:https://doi.org/10.1177/0022242921996646

[20] Jamie N. Mikeska, Calli Shekell, Jennifer Dix, and Pamela S. Lottero-Perdue. 2022. "Unnatural How Natural It Was": Using a Performance Task and Simulated Classroom for Preservice Secondary Teachers to Practice Engaging Student Avatars in Scientific Argumentation. *Journal of Technology and Teacher Education* 30, 3 (November 2022), 341–376. https://www.learntechlib.org/p/220407

[21] Shinji Miyake and Akinori Ito. 2012. A spoken dialogue system using virtual conversational agent with augmented reality. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. 1–4.

[22] Nooralisa Mohd Tuah, Vanissa Wanick, Ashok Ranchhod, and Gary Wills. 2017. Exploring avatar roles for motivational effects in gameful environments. *EAI Endorsed Transactions on Creative Technologies* 4 (09 2017), 153055. https://doi.org/10.4108/eai.4-9-2017.153055

[23] Heather L O'Brien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[24] ITU-T Recommendation P.851. 2003. Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems (11/2003). International Telecommunication Union.

[25] Dhaval Parmar, Joseph Isaac, Sabarish V Babu, Nikeetha D'Souza, Alison E Leonard, Sophie Jörg, Kara Gundersen, and Shaundra B Daily. 2016. Programming moves: Design and evaluation of applying embodied interaction in virtual environments to enhance computational thinking in middle school students. In *2016 IEEE Virtual Reality (VR)*. IEEE, 131–140.

[26] Mark Peterson. 2005. Learning interaction in an avatar-based virtual environment: a preliminary study. *PacCALL Journal* 1, 1 (2005), 29–40.

[27] Sihang Qiu, Alessandro Bozzon, Max V. Birk, and Ujwal Gadiraju. 2021. Using Worker Avatars to Improve Microtask Crowdsourcing. 5, CSCW2, Article 322 (oct 2021), 28 pages. https://doi.org/10.1145/3476063

[28] Niklas Rach, Klaus Weber, Annalena Aicher, Florian Lingenfelser, Elisabeth André, and Wolfgang Minker. 2019. Emotion recognition based preference modelling in argumentative dialogue systems. In *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 838–843.

[29] Rabindra Ratan, R.V. Rikard, Celina Wanek, Madison McKinley, Lee Johnson, and Young June Sah. 2016. Introducing Avatarification: An Experimental Examination of How Avatars Influence Student Motivation. https://doi.org/10.1109/HICSS.2016.15

[30] Genaro Rebolledo-Mendez, David Burden, and Sara de Freitas. 2008. A model of motivation for virtual-worlds avatars. *Lecture Notes in Computer Science* 5208 (2008), 535–536.

[31] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic Review: Trust-Building Factors and Implications for Conversational Agent Design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96. https://doi.org/10.1080/10447318.2020.1807710 arXiv:https://doi.org/10.1080/10447318.2020.1807710

[32] Mario Rincon-Nigro. 2013. A Text-Driven Conversational Avatar Interface for Instant Messaging on Mobile Devices. *Human-Machine Systems, IEEE Transactions on* 43 (05 2013), 328–332. https://doi.org/10.1109/TSMC.2013.2250498

[33] Roland Rust and P. K. Kannan. 2003. E-service: A new paradigm for business in the electronic environment. *Commun. ACM* 46 (06 2003), 37–42. https://doi.org/10.1145/777313.777336

[34] Ralph Schroeder. 2001. *The social life of avatars: Presence and interaction in shared virtual environments*. Springer Science & Business Media.

[35] Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays.. In *COLING*. 1501–1510.

[36] Ning Wang, W. Johnson, Richard Mayer, Paola Rizzo, Erin Shaw, and Heather Collins. 2005. The Politeness Effect: Pedagogical Agents and Learning Gains. 686–693.

[37] Janet Wessler, Tanja Schneeberger, Leon Christidis, and Patrick Gebhard. 2022. Virtual backlash: nonverbal expression of dominance leads to less liking of dominant female versus male agents. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*. 1–8.

[38] Thomas E Whalen, Dorina C Petriu, Lucy Yang, Emil M Petriu, and Marius D Cordea. 2003. Capturing behaviour for the use of avatars in virtual environments. *CyberPsychology & Behavior* 6, 5 (2003), 537–544.

[39] Catherine Zanbaka, Paula Goolkasian, and Larry Hodges. 2006. Can a Virtual Cat Persuade You? The Role of Gender and Realism in Speaker Persuasiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal, Québec, Canada) *(CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1153–1162. https://doi.org/10.1145/1124772.1124945