Aicher, Annalena; Weber, Klaus; André, Elisabeth; u. a.

BEA : Building Engaging Argumentation

Date of secondary publication: 15.10.2025

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-110816x

Primary publication

Aicher, Annalena; Weber, Klaus; André, Elisabeth; u. a. (2024): BEA : Building Engaging Argumentation, in: Philipp Cimiano, Anette Frank, Michael Kohlhase, u. a. (Ed.), Robust Argumentation Machines : First International Conference, RATIO 2024, Bielefeld, Germany, June 5–7, 2024, Proceedings, Cham: Springer Nature Switzerland, pp. 279–295, doi: 10.1007/978-3-031-63536-6_17.

BEA: Building Engaging Argumentation

Annalena Aicher[1,2](✉), Klaus Weber[2], Elisabeth André[2], Wolfgang Minker[1], and Stefan Ultes[3]

[1] Ulm University, Ulm, Germany
{annalena.aicher,wolfgang.minker}@uni-ulm.de
[2] University of Augsburg, Augsburg, Germany
{klaus.weber,elisabeth.andre,annalena.aicher}@uni-a.de
[3] University of Bamberg, Bamberg, Germany
stefan.ultes@uni-bamberg.de

Abstract. Exchanging arguments and knowledge in conversations is an intuitive way for humans to form opinions and reconcile opposing viewpoints. The vast amount of information available on the internet, often accessed through search engines, presents a considerable challenge. Managing and filtering this overwhelming wealth of data raises the potential for intellectual isolation. This can stem either from personalized searches that create "filter bubbles" by considering a user's history and preferences, or from the intrinsic, albeit unconscious, tendency of users to seek information that aligns with their existing beliefs, forming "self-imposed filter bubbles".

To address this issue, we introduce a model aimed at engaging the user in a critical examination of presented arguments and propose the use of a virtual agent engaging in a deliberative dialogue with human users to facilitate a fair and unbiased opinion formation. Our experiments have demonstrated the success of these models and their implementation. As a result, this work offers valuable insights for the design of future cooperative argumentative dialogue systems.

Keywords: Cooperative Argumentative Dialogue Systems · Reflective (User) Engagement (RUE) · Conversational Engagement · User Attention · User Focus · Gaze Tracking · Virtual Avatar

1 Introduction

Humans naturally form opinions and resolve differing perspectives through conversation, exchanging arguments and knowledge. Today's digital landscape offers

This work has been funded by the Deutsche Forschungsgemeinschaft within the project "BEA - Building Engaging Argumentation", Grant No. 455911629, as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).
A. Aiche and K. Weber—These authors contributed equally to this work and share first authorship. They are listed in alphabetical order.

P. Cimiano et al. (Eds.): RATIO 2024, LNAI 14638, pp. 279–295, 2024.
https://doi.org/10.1007/978-3-031-63536-6_17

a wealth of opinions and information available online anytime. However, navigating and evaluating this abundance of sources can be a challenging task. Filter algorithms attempt to alleviate this challenge by personalizing content based on users' past requests, potentially leading to the creation of so-called "filter bubbles" [23] where users are exposed to information that is mainly consistent with their existing viewpoints. In addition to these external influences, an intrinsically motivated counterpart comes to the forefront. Even when having access to non-filtered sources, people tend to prioritize a biased subset of sources that echo or reinforce their pre-existing or convenient opinions, forming so-called" **self-imposed** filter bubbles" (SFB) [4,7,10]. To counteract this unintentional intellectual isolation, we aim to 1) engage the user in an intuitive, fair and unbiased process of opinion formation, 2) enable the user to explore a wide range of information naturally and intuitively and thus, to "build(ing) engaging argumentation". To this end, we introduce the cooperative argumentative dialogue system BEA embodied by a virtual agent participating in a deliberative dialogue with a human user. Unlike persuasive systems with competitive agendas, our system aims to offer a diverse and representative overview within the context of a conversation with the user.

The primary goal of BEA is to establish an interactive platform that motivates users to explore diverse perspectives and critically scrutinize information on diverse topics. To overcome the limitations of a one-sided conversation, BEA leverages a flexible natural language understanding and multiple in- and output modalities. To provide a basis for a thorough, well-rounded discussion, we derive the necessary specific characteristics of the argumentative dialogue. These critical features include, first, the user's demonstration of critical thinking and open-mindedness during the interaction with the agent, the so-called *reflective engagement*. And second, the user's motivation in sustaining the conversation with the system represented by a human-like avatar, the so-called *conversational engagement*. The following paper aims to elucidate the architecture of the argumentative dialogue system BEA and its associated modules. In particular, it explains the underlying model for the user's reflective engagement and the corresponding intervention strategy of BEA based on this model. As an evaluation of our models and implementation in BEA, we present the results of two studies, demonstrating 1) increased user attention and focus on relevant parts of the arguments due to BEA's intervention and 2) the positive impact of a human-like virtual avatar embodying BEA on conversational user engagement, trust, and the general perception of the system.

The remainder of this paper is as follows: Sect. 2 provides a short overview of relevant related work. Section 3 gives an overview of the different components of BEA, such as the formal argument structure, dialogue framework, interface etc. In Sect. 4, we present an approach to model the user's reflective engagement. In Sect. 5 BEA's contribution in enhancing the reflective and conversational user engagement is evaluated. Respective limitations of our work are discussed in Sect. 6, followed by a conclusion and outlook on future work in Sect. 7.

2 Related Work

In the following, we give a short overview of the related work on 1) argumentative dialogue systems, 2) reflective engagement and 3) conversational user engagement and virtual Avatars.

2.1 Argumentative Dialog Systems

Argumentative dialogue systems (ADS), conversational agents (CA), and Chatbots aim to interact with users through natural language by exchanging arguments. Most approaches to human-machine argumentation are embedded in a competitive setting [27,28]. They utilize different models to structure the interaction (similarity model to retrieve counterarguments [25], retrieval- and generative-based models [16]). In contrast, [5] introduced a cooperative argumentative dialogue system that provides arguments upon users' request without trying to persuade or win a debate against the user. We adopt this cooperative approach, as a mere confrontation with opposing arguments leads to cognitive dissonance [13], which can have a negative effect (defensive attitude [12]). Therefore, a confrontation in a competitive scenario is more likely to lead to rejection.

2.2 Reflective Engagement

Reflective engagement (RE) in literature often denotes learners' active involvement in critically assessing their problem inquiry. Farr et al. [11] investigated markers of reflection in online discussions. Lyons et al. [17] emphasized the deliberate interruption of teaching practices for systematic questioning, highlighting the need for conscious awareness and adaptability. While existing research primarily explores RE in teaching-learning processes [14], our focus is on diverging viewpoints in argumentative scenarios. In contrast to methods like marker identification [11] our approach [30] integrates the user's stance and explored argumentation polarity for calculating reflective engagement, in line with [18]. Aicher et al. [3] introduced the reflective user engagement (RUE) score which aims for a balanced argument exploration of both sides. We extend this by rewarding users scrutinizing views opposing their current opinion. In their recent work [4,7] Aicher et al. introduced a model to determine the self-imposed filter bubble of the user and showed the effectiveness of the respective intervening "breaking strategy" to overcome the user's SFB [2].

2.3 Conversational User Engagement and Virtual Avatars

Current research often delves into the impact of self-identification with avatars in virtual spaces [20,26]). In contrast to virtual self-representations, our focus is on the influence of a virtual avatar as a discussion counterpart. Research in computer-mediated communication emphasizes that higher aesthetic and behavioral realism in avatars enhances user engagement, acceptance, and a sense of "social copresence." The results of Aseeri et al. [8] suggest that visual and non-verbal cues from different avatar representations affect user experience in cooperative tasks. In the context of argumentative dialogue systems, literature is very limited. Blount et al. [9] present an approach for participants to shape their own avatar appearance and how this impacts the course of an argumentative debate in the virtual sphere. However, there is a gap in analyzing the change in engagement, motivation, and perception when employing a virtual human-like avatar in a cooperative argumentative dialogue system instead of a chat-based interface, which we aim to address.

3 Prototype and Architecture of BEA

In the following, we give a short overview of prototype of BEA and its architecture, which is originally based on [5] and its extension [30].

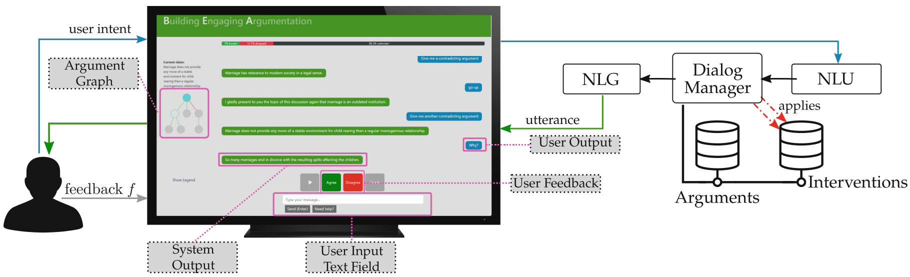

Fig. 1. Overview of system architecture in the interaction with a user.

3.1 System Architecture

Figure 1 sketches the architecture of our system. It consists of 1) an NLU (Natural Language Understanding), 2) a knowledge base of arguments, 3) a dialogue manager, 4) an NLG (Natural Language Generation), and 5) an intervention strategy.

Natural Language Understanding

The system uses an integrated natural language understanding framework (NLU) [1] to map the user's input to the available speech acts. The user can freely type their requests using a chat-input field to allow for a natural conversation. The NLU uses an intent classifier model consisting of two main components: a BERT Transformer Encoder and a bidirectional LSTM classifier.

Knowledge Base

The arguments that are available to the system throughout the interaction are encoded in an argument tree structure based on the argument annotation scheme introduced in [29]. The knowledge base consists of a set of components L_t. It includes three types of argument components (*Major Claim*, *Claim*, and *Premise*) and two different directed relations (*support* and *attack*) between them. Within the scope of this work, relations are allowed from *Claims* to the *Major Claim*, *Premises* to *Claims* and *Premises* to *Premises*. If a component $\varphi_i \in L_t$ has a relation towards a component $\varphi_j \in L_t$, we say that φ_j is the target (of φ_i) and each component (apart from the Major Claim φ_0) has exactly one target. Hence, the arguments $\Phi_i \in \boldsymbol{Args}$ that can be generated from such a structure have the form $\Phi_i = (\varphi_i \Rightarrow \varphi_j)$ ($\widehat{=}$ *support*) or $\Phi_i = (\varphi_i \Rightarrow \neg\varphi_j)$ ($\widehat{=}$ *attack*). Since each relation is unique and the difference between the three types of components is characterized solely by the allowed relations, each resulting structure can be represented as acyclic directed graph with argument components as nodes and relations as edges.

Throughout this work, we use the *idebate* dataset *Marriage is an outdated institution*[1] consisting of 72 arguments following the presented structure. The root argument is defined as $\Phi_0 := \varphi_0$. Every argument $\Phi_i \in Args$ has a stance $\in \{+, -\}$ towards φ_0 defined by the component relation and the respective position in the argument graph.

Table 1. Communication language L_c of the herein implemented dialogue system consisting of nine speech acts.

Speech Act	Description
System moves	
$argue(\varphi_i \Rightarrow \varphi_j)$	Present argument $\varphi_i \Rightarrow \varphi_j$
$jump\_to(\varphi_i)$	Jump to argument $\Phi_i = \varphi_i \Rightarrow *^{a}$
$intervene$	Suggest a challenger argument
User moves	
$why_{pro}(\varphi_i)$	Ask for a supporting component
$why_{con}(\varphi_i)$	Ask for an attacking component
$level_{up}$	Move level up
$agree(\varphi_i)$	Feedback to agree with a statement φ_i
$disagree(\varphi_i)$	Feedback to disagree with a statement φ_i
$confirm/reject$	Confirm/Reject intervention

[a] $* \in \{\varphi, \neg\varphi\}$

Dialogue Manager

The dialogue manager has access to different knowledge bases and provides a communication language L_c, which includes the speech acts available to the

[1] https://idebate.net/resources/debatabase.

user and system (see Table 1). These speech acts are tailored to suit the purpose of a specific dialogue system and can be modified accordingly. It manages the dialogue between the system and the user and ensures logical consistency. Furthermore, the dialogue manager stores the current dialogue state, i.e., the complete dialogue history, which arguments have been presented, the current position within the argument tree, and allowed speech acts. For instance, if an argument Φ_i is a leaf node, $why_{pro}(\varphi_i)$ is not allowed. If the user requests a new argument (why_x), the system selects a random argument from all arguments fitting the requested relation $x \in \{pro, con\}$.

Natural Language Generation
The system generates a textual response, wherein the Natural Language Generation (NLG) relies on the original surface text of the argument components, denoted as $\varphi_i \in L_t$. These annotated sentences were manually adjusted in terms of grammatical syntax to create independent utterances, serving as templates for the corresponding system responses. In order to add some diversity, a collection of natural language formulations was created for each speech act. During the response generation, the explicit formulation is chosen from this list randomly. To structure the dialogue as comprehensible and understandable for the user as possible, and to clearly present contextual connections, the system employs transitional phrases such as *Let us return to the previous argument, that ...* or *This claim is supported by the argument that ...*. Please note that the system presents all arguments in a neutral manner, without taking a stance of its own.

Intervention Strategy
The intervention keeps track of the user's reflective engagement (RUE) (RUE, see Sect. 4 for calculations) and intervenes if necessary, i.e., it suggests considering an opposing viewpoint and presents a challenger argument if the user agrees. Let $l^+, l^- \in L_t$ be the set of all valid[2] argument components that the system can present. For every user request, the system simulates the RUE_{Φ_0} for all valid challenger arguments and returns the component that maximizes it. The intervention takes place if and only if the simulated RUE is greater than the maximum possible RUE for all available non-challenger arguments, i.e., if the user requests a supporting argument, the system checks if

$$\max_{\varphi_i \in l^-}(sim\_rue(\varphi_i)) > \max_{\varphi_i \in l^+}(sim\_rue(\varphi_i)) \tag{1}$$

and, if so, returns $\varphi_i = \arg\max_{\varphi_i \in l^-}(sim\_rue(\varphi_i))$, and suggests it to the user, who can accept or reject it. In case of rejection, the system proceeds with the initial user request.

3.2 User Interface

Table 2 sketches a sample dialogue between the user and system.

[2] A valid pro (+) and con (-) argument component φ_i is an unheard component that links to any component φ_j that has already been presented to the user.

Table 2. Exemplary dialogue between the system and the user about the topic *"Marriage is an outdated institution"* with intervention condition.

Actor	Utterance/Action	Speech Acts
system	*Let's discuss that marriage is an outdated institution.*	claim(φ_0)
User	Give me an argument in favor.	$why_{pro}(\varphi_0)$
system	*Marriage does not provide a more stable environment for child rearing than a regular monogamous relationship.*	$argue(\varphi_1 \Rightarrow \varphi_0)$
User	*Clicks green button "Agree" on interface.*	$agree(\varphi_1)$
system	*Alright, I noted your feedback. Let us return to the topic of this discussion.*	-
User	Could you elaborate this further?	$why_{pro}(\varphi_0)$
system	*I think we should look at the opposite point of view. Alright?*	*intervene*
User	Yes	$confirm$
system	*Marriage removes the transient and casual aspects of a monogamous relationship, thus giving a child a far more stable environment.*	$argue(\varphi_2 \Rightarrow \neg\varphi_0)$

The user interface consists of four components: 1) the dialogue history showing the system output and user input, 2) the user input text field, 3) the user feedback buttons, and 4) the graphically displayed information about the argument structure. Furthermore, instead of the dialogue history in the middle of the screen, a human-like 3D avatar can be displayed in the same position, presenting the arguments in spoken language (see Fig. 4). The dialogue history is depicted on the right side of the screen in this configuration.

A classic chat design displays the system's *textual* response. The interface provides a *text* (chat) input where users can formulate their requests. To allow for feedback on whether users agree with the current argument or not, there are two buttons (*Agree* and *Disagree*) that the user can use at any time during interaction. Without feedback, the system considers a neutral user stance. On the left side the browser window left side, the current argument graph displays a visual representation of all arguments of the respective *Root Claim*[3]. An outlined turquoise node denotes the user's current position, the already discussed arguments are shown in solid turquoise, and unheard arguments are in grey. The edges between the nodes show a supporting relation in green and an attacking one in red. Users can choose whether to ask for a pro or con argument or how they want to navigate through the argument tree.

[3] We define a *Root Claim* as a *Claim* φ_j which directly *attacks* or *supports* the *Major Claim* φ_0.

4 Modeling Reflective Engagement

In the following, we give a short overview of the reflective engagement model (for details see [3,30]).

A set of arguments with the same target argument Φ_i is denoted as P_{Φ_i}. If it is in favor of stance +, it is denoted as $P^+_{\Phi_i}$ and $P^-_{\Phi_i}$ elsewise. The set of all visited arguments Φ_j with target argument Φ_i is denoted as $P_{\Phi_i,v}$.

We define the user's focus for argument Φ_i based on visited pro and con arguments as:

$$focus_{\Phi_i} = \frac{\left|P^+_{\Phi_i,v}\right| - \left|P^-_{\Phi_i,v}\right|}{\left|P_{\Phi_i,v}\right|} \in [-1, 1]. \tag{2}$$

It is easy to verify that the more arguments of a certain stance are selected by the user, the more the focus shifts in the direction of the respective stance.

The overall normalized user focus $\mathcal{F} \in [-1, 1]$ is then defined by summing up and normalizing the $focus_{\Phi_i}$:

$$\mathcal{F} := \frac{\sum_{\Phi_k \in Args} focus_{\Phi_k}}{|Args|} \tag{3}$$

During the interaction with the system, users can give feedback on whether or not they agree or disagree with any argument $\Phi_i \in Args$. Considering the hierarchical structure of arguments, the system uses this feedback to compute the user's stance $e_{\Phi_j} \in [0, 1]$ of any argument $\Phi_j \in Args$ considering the feedback for this respective argument and the feedback for all arguments in the subtree with Φ_j as root (following the approach of [31]).

The user's reflective engagement considers the weighted user's focus by making use of the inverted correlation of stance and focus (Eq. 4). This is based on the assumption that users with a particular stance are likely to focus more on arguments that are in line with their stance. Users with a higher level of RE tend to look at claims that support an opposite view as well [24].

$$RUE = 1 - \left| e_{\Phi_0} - \left(1 - \frac{\mathcal{F} + 1}{2}\right)\right| \tag{4}$$

After inverting the normalized focus, the difference between user stance and focus is taken to compute RUE, i.e., the more the focus aligns with the user's stance, the lower the RUE and vice versa. This approach ensures that if the user stance is positive (+), the system intervenes to suggest the user choose more con arguments (challenger arguments of pro arguments) and vice versa.

5 Evaluation

In the following we give an overview of the results of to our two previously published evaluation studies[4], aimed to analyze the influence of 1) the intervention

[4] Due to space limitations, only essential results are presented here; further details are available in our cited work.

on the user's focus on challenger arguments [30] and 2) a virtual human-like avatar on the general user perception and conversational user engagement [6].

5.1 Study 1 [30]: Analyzing Focus on Challenger Arguments

The first study was conducted online via the crowdsourcing platform "Crowdee[5]" with 58 participants (aged 18–63) divided into two groups (an experimental group with intervention and a control group without intervention) from the UK, US, and Australia (English native speakers to avoid language barrier effects). The study setup used the chat-based output modality. After an introduction to the system (short text and description of how to interact with the system), the users were advised 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 underlying reflection model but only to select at least ten arguments. In addition, they were asked to rate their opinion on the topic on a 5-point Likert scale, which normalized in $[0, 1]$ displayed the initial user stance e_0. During the study, we collected the following data anonymously:

1. Measured user reflection score RUE (Fig. 2a).
2. User stance e_{Φ_0} (Fig. 2b)
3. Set of visited arguments P_v^+ and P_v^- (Fig. 2c).

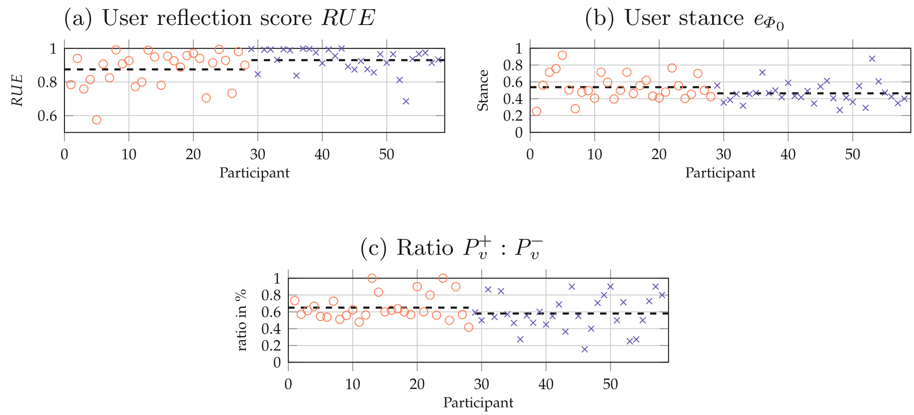

Fig. 2. Collected data: Reflection score RUE, user stance e_{Φ_0}, and focus $\mathcal{F}$ of both the experimental group (x) and the control group (o).

Statistical Analysis: Concerning the calculated RUE score (Fig. 2a), the homogeneity of variances was falsified utilizing the Levene's test ($F = 5.64$, $p = .021$) and the assumption of a normal distribution using the Shapiro-Wilk

[5] https://www.crowdee.com/.

test ($W = 0.895$, $p < .001, W = 0.874$). Thus, we applied the *Mann-Whitney-U test* showing a main effect of intervention on RUE ($U = 273$, $n_1 = 30$, $n_2 = 28$, $p \leq .01$). In addition to that, we also checked the total amount of interventions. There were 262 interventions in the experimental condition (8.73 per user), 201 of which were accepted by the user, which is an acceptance rate of 76%.

To investigate the main effect of intervention on challenger arguments, we analyzed how many participants were more engaged with *challenger arguments* by comparing the amount of pro and con arguments that the user heard to the user stance, e.g., if the user stance is negative ($e_{\Phi_0} < 0.5$) and more pro than con arguments were heard ($P_v^+ > P_v^-$), it implicates a strong challenged engagement (see Table 3).

Table 3. Contingency table of focus on challenger arguments per condition.

Condition	Challenger arg.	Non-challenger arg.	Total
Experimental	24	6	30
Control	15	13	28
Marginal Column Total	39	19	58

We found that with intervention, nearly 80% of participants were more engaged with *challenger arguments*, while only 53% in the control condition did so, which is a total increase of 51%. A chi-square test of independence [19] was performed to examine the relation between *condition* and *engagement with challenger arguments* showing a significant relation ($\chi^2(1, N = 58) = 4.5924, p = .032$).

Analyzing the main effect of intervention on the total percentage of heard *challenger arguments* revealed a large significant main effect (*T-Test*, $t(56) = 2.0903, p = .02, d = .55$, Fig. 3), proving that the users in the experimental group focused significantly more on *challenger arguments* than in the control condition.

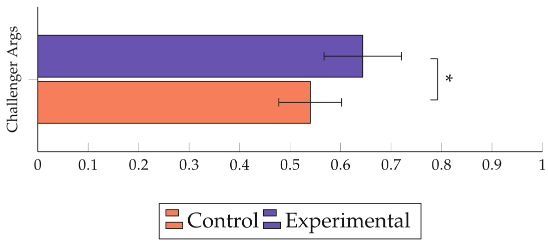

Fig. 3. Means including 95% confidence interval denoted by bars of focus on challenger arguments. (*) $p < .05$.

5.2 Study 2 [6]: Influence of Avatar Interface

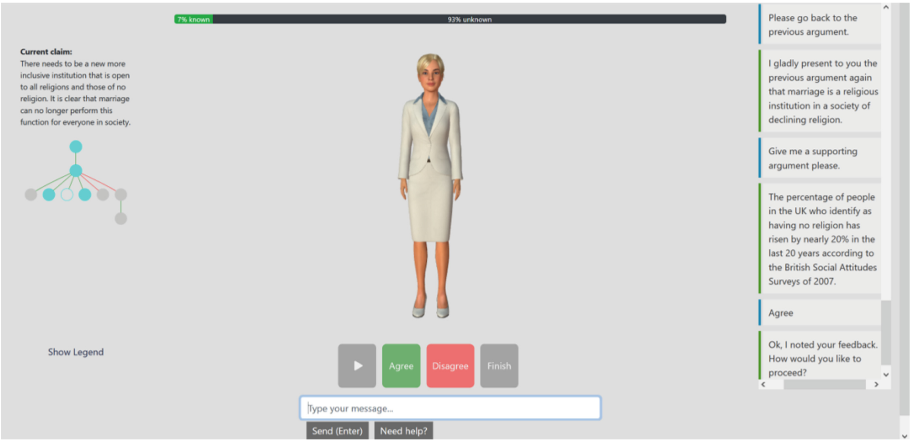

Fig. 4. User interface with avatar. Above the chat-input line four buttons and the virtual avatar are shown. The dialogue history is placed on the right side of the screen.

After having shown the positive impact of BEA's intervention on the user's focus and reflective engagement, this subsection focuses on analysing how to create a enjoyable, natural discussion and maintain the user motivation to interact with BEA. Therefore, we examine the impact of avatar versus non-avatar interfaces on user perception, engagement, and trust in argumentative dialogue systems, we conducted a crowdsourcing study. Eighty-four participants (aged 18–65; 52 female, 31 male, 1 "other/do not want to say") were divided into two groups: 46 interacted with a virtual avatar interface (avatar system) illustrated in Fig. 4, and 38 with a non-avatar interface (non-avatar system) (similar to the screen in Fig. 1). Both systems were identical, differing only in the graphical user interface (chat-based output for the non-avatar system, spoken avatar output for the avatar system). The avatar interface utilized the Charamel^{TM} avatar[6] with synthetic speech utilizing Nuance TTS and Amazon Polly Voices[7].

Participants interacted significantly longer with the avatar, influenced by the avatar's spoken utterance and response delays caused by the avatar server. The participants rated statements on a 5-point Likert scale (1="Totally disagree", 5="Totally agree") across three questionnaires. In addition to the three questionnaires, we asked the participants to rate their opinion and interest in the topic "Marriage is an outdated institution" before ("pre") and after ("post") the interaction. There is no significant difference between the two participant groups (*Mann-Whitney-U-Test*) or between the pre- and post-conditions (*Wilcoxon*

[6] https://www.charamel.com/competence/avatare.
[7] https://docs.aws.amazon.com/polly/latest/dg/voicelist.html.

signed-rank test) regarding user opinion. This is important to avoid the risk that the avatar itself biases and manipulates user opinion. While the differences in user interest are not yet significant, they are still noticeable for the avatar group ($Mean_{pre} = 3.56$, $Mean_{post} = 3.81$,$p = .074$, $r = .195$) compared to the non-avatar one ($Mean_{pre} = 3.58$, $Mean_{post} = 3.53$,$p = .614$, $r = .055$). This implies a slight tendency that the user interest is positively influenced by the avatar.

First questionnaire (adopted from a questionnaire according to ITU-T Recommendation P.851) [22][8]: consists of 39 single items and measures the user's general impression of the system. Its items are grouped by the following aspects: information provided by the system (IPS), communication with the system (COM), system behavior (SB), dialogue (DI), user's impression of the system (UIS), acceptability (ACC), and argumentation (ARG)[9].

Even though the differences between the avatar and non-avatar group regarding the merged aspects IPS, COM, SB, DI, UIS, ACC and ARG are insignificant (*Mann-Whitney-U-Test*), we can perceive some consistent, aspect-overlapping tendency. Regarding the aspects IPS, COM, UIS and ACC neither the single item analysis nor the merged analysis showed any significant differences. Likewise also for the three other aspects (SB, DI and ARG) the merged analysis did not show any significant differences, but we could still perceive some consistent, aspect-overlapping tendency. Especially regarding the perceived naturalness and the engagement users felt, the avatar system is rated significantly better.

Second questionnaire: consists of 12 items [21] and measures the conversational engagement. Its items are grouped by the following aspects: Focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and Reward (RW). The findings revealed the avatar system's engaging effect, as indicated by better ratings across all items and especially significant for the impression that using the system was worthwhile ($p_{RW1} = .022, r_{RW1} = 0.25$). However, perceived usability needs improvement, particularly regarding automated speech recognition errors and the explanation of the system's reaction if the user was not understood correctly.

Together with the voluntary the significant difference in the expected help the system should have provided, this implies that the avatar on one hand side tends to raise the expectation to that of a human conversational partner. Thus, fulfilling these expectations could lead to a significantly stronger acceptance comparable to a human conversational partner.

Third questionnaire [15][10]**:** consists of 11 items and measures user trust. Its items are grouped by the following aspects: understanding/predictability (UP), familiarity (F), propensity to trust (PT) and trust in automation (TA). The users tend to trust the avatar system more than the non-avatar one, especially regarding their propensity to trust ($p_{PT} = 0.015$, $r_{PT} = 0.266$). Both participant

[8] Such questionnaires can be used to evaluate the quality of speech-based services.
[9] Self-added aspect since this is not captured by standardized questionnaires.
[10] This questionnaire was developed to measure trust in automation.

groups do not differ noticeably in the familiarity with similar systems, which implies an inclination towards trusting the avatar system more. Thus, especially by individualizing the avatar further, we believe to increase the user trust and support a well-founded opinion building.

In summary, our findings support using an avatar interface in ADS, emphasizing its potential to enhance user engagement and trust without manipulating their opinion. Addressing usability concerns can further optimize the user satisfaction and it will become easier to maintain the interaction.

6 Limitations

This paper, however, is subject to some limitations that will be addressed in future research. First, in our second study, we did not compare different avatar settings personalized to individual users, nor did we conduct per-participant analyses. For this exploratory study, we opted for an easily implementable, widely accessible, representative avatar, rather than one that is highly individualized. The comparison to a purely chat-based interface aimed to evaluate the influence of avatars on argumentative interactions in general. The focus was to determine whether the mere visualization of an avatar leads to a bias in opinion formation or influences the perception of the provided argumentative content and conversational engagement of users even though the avatar has not been personalized. However, future research in this domain should investigate the impact of personalized avatar features and how individual participants perceive the interaction, the provided argument content, and the avatar's personality traits (e.g., dominance, friendliness/pleasure). In order to capture the full range of experiences and perspectives of the participants in future studies, it may be beneficial to supplement the study with qualitative data in the form of participant interviews (e.g. by free text responses). Furthermore, it needs to be mentioned, that due to limited space we did not discuss user comments in detail. Moreover in future work we will put a specific focus on analyzing aspects directly related to argumentation, such as the perception of argument strategy/selection (regarding consistency, quality, persuasiveness, etc.) in relation to differences in avatar modeling.

Another limitation is that both user studies focus solely on one topic ("Marriage is an outdated institution") derived from a single source. We selected this topic because its dataset fulfills our criteria of being sufficiently large, balanced in terms of argument stance (pro/con), of high quality, and having depth in arguments. Although it appears suitable for a proof-of-principle study, the scalability of our findings needs to be demonstrated concerning other topics.

Moreover, we emphasize that while the user-agent interaction may seem constrained and artificial because users are unable to introduce counterarguments, this decision was deliberate. The aim of the argumentative dialogue system is to neutrally confront users with pro/con arguments on a given topic, allowing them to explore without being directly engaged in a persuasive discussion. As pointed out by [24] due to the users' tendency to defend their own view, a system which confronts them with an opposing stance might not lead to critical reflection but rather the opposite. However, we aim for a more natural exchange in future work. We intend to explore how users can introduce their own arguments without exacerbating their self-imposed filter bubble. To achieve this, we propose investigating the approach of dynamically searching for relevant arguments in real-time, which can support both viewpoints and supplement the existing argumentative structure with arguments that have not yet been part of the argumentative discourse.

One final limitation to mention here is that the study results presented herein have predominantly been introduced separately in our earlier publications. The intention behind this is to integrate and merge findings from previously separately explored dimensions of reflective and conversational engagement within argumentative discourse involving real users. This synthesis shall provide a basis for addressing questions that involve both forms of engagement, conversational and reflective engagement, to enhance argumentative discussions with individual users in future work. Thus in this paper, we want to emphasize the importance of acknowledging the significance of investigating both conversational and reflective engagement to attain a critically-reflected and enjoyable interaction with real users within the context of robust argumentation machines.

7 Conclusion and Future Work

In this work we introduced an approach to build engaging argumentation. Therefore, we focused on exploring methods to enhance the reflective and conversational engagement of users in the interaction with the cooperative argumentative dialogue system BEA. To account for a critical reflection of arguments, we introduce a model that characterizes reflective user engagement (RUE) and propose a corresponding intervention strategy. To maintain the user motivation to continue the interaction, especially when guiding them to explore opposing viewpoints without explicit request, we suggest incorporating a virtual, human-like avatar to embody the system.

The first of two presented user studies demonstrates that BEA's intervention not only increased RUE but also enhanced the user focus on challenger arguments. Results from a second user study imply the positive impact of a human-like virtual avatar embodying BEA on conversational user engagement, trust, and overall system perception. This suggests that a human-like design generates expectations for communication and assistance, akin to interactions with a human conversational partner without manipulating the user's opinion. However, errors or delays in the avatar's response time can have a noticeable adverse impact, emphasizing the need for addressing these issues in future refinements of the avatar setting.

Also, in future work, we aim to explore how personalizing and individualizing avatars influence users' perception of argument content (persuasiveness, etc.) and user trust and motivation, particularly in relation to specific avatar traits. Through qualitative analyses, we hope to gain more insight into the factors that contribute to manipulating the rational perception of argument content when modeling virtual, argumentative avatars. And we will investigate the potential of personalized avatar implementation to enhance interaction in argumentative dialogue systems, through social presence influence, positive feedback, and emotional connection, with the intention to increased user engagement and satisfaction.

To ensure scalability of the described results and introduced approaches, we aim to test them on datasets covering other controversial topics from various sources. Additionally, we aim to enhance natural interaction by allowing users to contribute their own aspects/arguments while ensuring that the user's self-imposed filter bubble does not reinforce.

In conclusion, based on the established connection between reflective and conversational engagement in argumentative interaction with real users, we aim to further explore the interdependencies between them and their impacts. Therefore, this work presents important implications for designing a critically-reflected and enjoyable interaction within the context of robust argumentation machines.

References

1. Abro, W.A., Aicher, A., Rach, N., Ultes, S., Minker, W., Qi, G.: Natural language understanding for argumentative dialogue systems in the opinion building domain. Knowl. Based Syst. **242**, 108318 (2022). https://www.sciencedirect.com/science/article/pii/S0950705122001149
2. Aicher, A., Kornmueller, D., Matsuda, Y., Ultes, S., Minker, W., Yasumoto, K.: Towards breaking the self-imposed filter bubble in argumentative dialogues. In: Stoyanchev, S., Joty, S., Schlangen, D., Dusek, O., Kennington, C., Alikhani, M. (eds.) Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pp. 593–604, September 2023. https://aclanthology.org/2023.sigdial-1.56
3. Aicher, A., Minker, W., Ultes, S.: Determination of Reflective User Engagement in Argumentative Dialogue Systems (2021)
4. Aicher, A., Minker, W., Ultes, S.: Towards modelling self-imposed filter bubbles in argumentative dialogue systems. In: Proceedings of the 13th LREC, pp. 4126–4134, June 2022. https://aclanthology.org/2022.lrec-1.438
5. Aicher, A., Rach, N., Minker, W., Ultes, S.: Opinion building based on the argumentative dialogue system BEA. In: Marchi, E., Siniscalchi, S.M., Cumani, S., Salerno, V.M., Li, H. (eds.) Increasing Naturalness and Flexibility in Spoken Dialogue Interaction. LNEE, vol. 714, pp. 307–318. Springer, Singapore (2021). https://doi.org/10.1007/978-981-15-9323-9_27
6. Aicher, A., Weber, K., André, E., Minker, W., Ultes, S.: The influence of avatar interfaces on argumentative dialogues. In: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents. IVA 2023 (2023).https://doi.org/10.1145/3570945.3607343
7. Aicher, A.B., Kornmüller, D., Minker, W., Ultes, S.: Self-imposed filter bubble model for argumentative dialogues. In: Proceedings of the 5th International Conference on Conversational User Interfaces. CUI 2023 (2023).https://doi.org/10.1145/3571884.3597131

8. Aseeri, S., Interrante, V.: The influence of avatar representation on interpersonal communication in virtual social environments. IEEE Trans. Visual Comput. Graph. **27**(5), 2608–2617 (2021). https://doi.org/10.1109/TVCG.2021.3067783
9. Blount, T., Millard, D.E., Weal, M.J.: On the role of avatars in argumentation. In: Proceedings of the 2015 Workshop on Narrative & Hypertext, pp. 17–19. NHT 2015, Association for Computing Machinery (2015).https://doi.org/10.1145/2804565.2804569
10. Ekström, A.G., Niehorster, D.C., Olsson, E.J.: Self-imposed filter bubbles: selective attention and exposure in online search. Comput. Human Behav. Rep. **7**, 100226 (2022). https://doi.org/10.1016/j.chbr.2022.100226
11. Farr, F., Riordan, E.: Students' engagement in reflective tasks: an investigation of interactive and non-interactive discourse corpora. Classroom Disc. **3**(2), 129–146 (2012). https://doi.org/10.1080/19463014.2012.716622
12. Harmon-Jones, E.: Cognitive dissonance and experienced negative affect: evidence that dissonance increases experienced negative affect even in the absence of aversive consequences. Pers. Soc. Psychol. Bull. **26**(12), 1490–1501 (2000). https://doi.org/10.1177/01461672002612004
13. Hart, W., Albarracín, D., Eagly, A.H., Brechan, I., Lindberg, M.J., Merrill, L.: Feeling validated versus being correct: a meta-analysis of selective exposure to information. Psychol. Bull. **135**(4), 555 (2009)
14. Kong, S.C., Song, Y.: An experience of personalized learning hub initiative embedding BYOD for reflective engagement in higher education. Comput. Educ. **88**, 227–240 (2015). https://doi.org/10.1016/j.compedu.2015.06.003
15. Körber, M.: Theoretical considerations and development of a questionnaire to measure trust in automation. In: Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., Fujita, Y. (eds.) IEA 2018. AISC, vol. 823, pp. 13–30. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-96074-6_2
16. Le, D.T., Nguyen, C.T., Nguyen, K.A.: Dave the debater: a retrieval-based and generative argumentative dialogue agent. In: Proceedings of the 5th Workshop on Argument Mining, pp. 121–130 (2018).https://doi.org/10.18653/v1/W18-5215
17. Lyons, N.: Reflective engagement as professional development in the lives of university teachers. Teach. Teach. **12**(2), 151–168 (2006)
18. Mason, M.: Critical thinking and learning. Educ. Philos. Theory **39**(4), 339–349 (2007). https://doi.org/10.1111/j.1469-5812.2007.00343.x
19. McHugh, M.L.: The chi-square test of independence. Biochemia medica **23**(2), 143–149 (2013).https://doi.org/10.11613/BM.2013.018
20. Mohd Tuah, N., Wanick, V., Ranchhod, A., Wills, G.: Exploring avatar roles for motivational effects in gameful environments. EAI Endorsed Trans. Creat. Tech. **4**, 153055 (2017).https://doi.org/10.4108/eai.4-9-2017.153055
21. O'Brien, H.L., Cairns, P., Hall, M.: A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. Int. J. Human-Comput. Stud. **112**, 28–39 (2018)
22. P.851, I.T.R.: Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003). Int. Telecommunication Union, November 2003
23. Pariser, E.: The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Books, USA (2012)
24. Paul, R.W.: Critical and reflective thinking: A philosophical perspective, pp. 445–494 (1990), north Central Regional USA
25. Rakshit, G., Bowden, K., Reed, L., Misra, A., Walker, M.: Debbie, the debate bot of the future. In: IWSDS 2017, pp. 45–52, June 2017

26. Ratan, R., Rikard, R., Wanek, C., McKinley, M., Johnson, L., Sah, Y.J.: Introducing avatarification: an experimental examination of how avatars influence student motivation (2016).https://doi.org/10.1109/HICSS.2016.15
27. Rosenfeld, A., Kraus, S.: Strategical argumentative agent for human persuasion. In: ECAI 2016, pp. 320–328 (2016).https://doi.org/10.3233/978-1-61499-672-9-320
28. Slonim, N., et al.: An autonomous debating system. Nature **591**(7850), 379–384 (2021). https://doi.org/10.1038/s41586-021-03215-w
29. Stab, C., Gurevych, I.: Annotating argument components and relations in persuasive essays. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics Technical Papers, pp. 1501–1510, August 2014. https://aclanthology.org/C14-1142
30. Weber, K., Aicher, A., Minker, W., Ultes, S., André, E.: Fostering user engagement in the critical reflection of arguments. In: Proceedings of the 13th International Workshop On Spoken Dialogue Systems (IWSDS) (2023), accepted for publication
31. Weber, K., et al.: Predicting persuasive effectiveness for multimodal behavior adaptation using bipolar weighted argument graphs, pp. 1476–1484 (2020)