Multimodal Technologies and Interaction

MDPI

Article

Enhancing Reflective and Conversational User Engagement in Argumentative Dialogues with Virtual Agents

Annalena Aicher [1,2,3,*], Yuki Matsuda [2], Keichii Yasumoto [2], Wolfgang Minker [1], Elisabeth André [3] and Stefan Ultes [4]

[1] Institute of Communications Engineering, Ulm University, 89081 Ulm, Germany
[2] Ubiquitous Computing Systems Laboratory, Nara Institute of Science and Technology, Ikoma 630-0192, Nara, Japan
[3] Human-Centered Artificial Intelligence, University of Augsburg, 86159 Augsburg, Germany
[4] Natural Language Generation and Dialogue Systems, University of Bamberg, 96050 Bamberg, Germany
* Correspondence: annalena.aicher@uni-ulm.de

Abstract: In their process of information seeking, human users tend to selectively ignore information that contradicts their pre-existing beliefs or opinions. These so-called "self-imposed filter bubbles" (SFBs) pose a significant challenge for argumentative conversational agents aiming to facilitate critical, unbiased opinion formation on controversial topics. With the ultimate goal of developing a system that helps users break their self-imposed filter bubbles (SFBs), this paper aims to investigate the role of co-speech gestures, specifically examining how these gestures significantly contribute to achieving this objective. This paper extends current research by examining methods to engage users in cooperative discussions with a virtual human-like agent, encouraging a deep reflection on arguments to disrupt SFBs. Specifically, we investigate the agent's non-verbal behavior in the form of co-speech gestures. We analyze whether co-speech gestures, depending on the conveyed information, enhance motivation, and thus conversational user engagement, thereby encouraging users to consider information that could potentially disrupt their SFBs. The findings of a laboratory study with 56 participants highlight the importance of non-verbal agent behaviors, such as co-speech gestures, in improving users' perceptions of the interaction and the conveyed content. This effect is particularly notable when the content aims to challenge the user's SFB. Therefore, this research offers valuable insights into enhancing user engagement in the design of multimodal interactions with future cooperative argumentative virtual agents.

Keywords: non-verbal agent behavior; virtual agents; co-speech gestures; user interest; confirmation bias; self-imposed filter bubbles (SFB); argumentative dialogue systems (ADS); human–agent interaction

Citation: Aicher, A.; Matsuda, Y.; Yasumoto, K.; Minker, W.; André, E.; Ultes, S. Enhancing Reflective and Conversational User Engagement in Argumentative Dialogues with Virtual Agents. *Multimodal Technol. Interact.* **2024**, *8*, 71. https://doi.org/10.3390/mti8080071

Academic Editors: Gerd Bruder and Mu-Chun Su

Received: 30 March 2024
Revised: 19 July 2024
Accepted: 19 July 2024
Published: 6 August 2024

1. Introduction

Language, voice, posture, gestures, and gaze are examples of multimodal interaction, vital for human social communication. Effective communication integrates verbal and non-verbal signals, where gestures and facial expressions play a key role in conveying ideas [1]. As virtual assistants and agents become more prevalent, there is a growing emphasis on making their interactions more natural. To achieve this, non-verbal cues must be perceived, understood, and generated in a multimodal manner. In this paper, we focus on co-speech gestures, defined as "motions and poses primarily made with the arms and hands (or the upper body parts) in order to communicate while speaking" [2], in the context of interaction with a virtual agent embodying an argumentative dialogue system.

Spoken dialogue systems facilitate users' access to information from online sources, like search engines or social networks. However, particularly in complex interactions, Nickerson [3] claims that users tend to focus on a "biased subset of sources that repeat or strengthen an already established or convenient opinion", leading to "self-imposed filter bubbles"(SFB) [4,5]. SFBs differ from "filter bubbles" and "echo chambers" [6–8], which,

although similar, arise from different causes. According to Pariser [9], filter algorithms create cultural/ideological bubbles based on previous online behavior, while echo chambers describe environments that expose a person only to information or opinions that reflect and reinforce their own. Being trapped in a self-imposed filter bubble results from selective engagement with information in search results [4] and disregarding contradictory information [10].

While numerous empirical studies [11,12] suggest that echo chambers may be smaller than commonly believed and refute the filter bubble hypothesis, it is important not to adopt a Panglossian perspective that ignores the serious societal challenges posed by our digital, mobile, and platform-dominated media landscape. According to [13], these challenges include pronounced inequality in news consumption, widespread online harassment, and misinformation, which underscores the importance of developing systems that support users in breaking their SFBs.

In previous work [5,14], we defined four main dimensions of an SFB: Reflective User Engagement (RUE) [15], Personal Relevance (PR), True Knowledge (TK), and False Knowledge (FK) to model the user's SFB, along with a strategy to "break" and overcome this SFB [16].

To move closer to our overarching goal of developing a system that helps users break their SFBs, this paper aims to investigate the role of co-speech gestures, specifically examining how these gestures significantly contribute to achieving this objective. This paper extends current research by examining methods to engage users in cooperative discussions with a virtual human-like agent, encouraging a deep reflection on arguments to disrupt SFBs. Specifically, we investigate the agent's non-verbal behavior in the form of co-speech gestures. The results of our previous research [17] showed that the use of an avatar positively influences user engagement, trust, and perception of naturalness in interaction, with a human-like design raising expectations of communication and help, similar to a human conversational partner. Building on these findings, we analyze the extent to which non-individualized co-speech gestures of the agent have a positive effect on the user's engagement, the general perception of the interaction with the agent, and the provided content, especially when the agent simultaneously attempts to break the user's SFB.

So far, within the argumentation domain, the existing literature remains fragmented [18]. The specific impact of virtual agents and their behavior on the dynamics of argumentative debates remains unclear [19]. This paper aims to address this gap by investigating the influence of co-speech gestures of a human-like virtual agent in cooperative argumentative dialogues when specific argument selection strategies are applied.

The main contributions of this article are as follows:

1. Introduction to our self-imposed filter bubble model [5,14] and the respective SFB-breaking strategy [16] for selecting suitable content (arguments) in the interaction.
2. Introduction to our user Interest Model [20] and the corresponding interest (I) strategy for selecting arguments that fit the user's interests best.
3. Analyzing the influence of the applied argument selection strategy (SFB-breaking, I) on the user's overall agent perception (For simplicity, "overall agent perception" is used as the abbreviation for "the user's overall perception of the virtual agent representing the argumentative cooperative dialogue system".), the user's engagement, and opinion formation.
4. Analyzing the influence of non-verbal agent behavior (co-speech gestures) on the user's overall agent perception, the user's engagement, and opinion formation with respect to the provided content chosen by the two selection strategies (SFB-breaking, I).

This paper builds upon and extends work published previously [5,16,20,21]. Whereas in [21] we addressed the general influence of co-speech gestures, within this paper, we focus on how the influence of the agent's co-speech gestures becomes noticeable depending on different argumentative content chosen according to two different argument selection strategies (Interest, SFB-breaking). In particular, we want to investigate whether non-

verbal agent behavior in the form of co-speech gestures can mitigate the potential negative effects of the SFB-breaking selection strategy compared to an argument selection that best fits the user's interest. Therefore, we explore methods to enhance the user's reflective engagement by selecting respective SFB-breaking arguments and the user's conversational engagement by using co-speech gestures of the embodied conversational agent (ECA) during the interaction.

The structure of the paper is as follows: In Section 2, we focus on the relevant literature on both, reflective and conversational engagement in the context of argumentative dialogue systems and provide the background for the subsequent section. In this Section 3, we introduce the Interest Model, the SFB Model and the respective argument selection strategies. Section 4 describes the architecture of the argumentative dialogue system (ADS), especially with respect to the dialogue model and interface with non-verbal agent behavior, i.e., co-speech gesture modeling. Afterward, Section 5 covers the experiment and study setup before describing the evaluation results in Section 6. The discussion of the results and limitations is included in Sections 7 and 8 followed by a conclusion and outlook on future work in Section 9.

2. Reflective and Conversational Engagement in the Interaction with ECA

In this section, we first examine the literature on reflective engagement and then on conversational engagement, focusing on argumentative embodied conversational agents.

Here, reflective engagement refers to a cognitive process in which an individual actively thinks about, analyzes, and evaluates information or experiences. It involves deep introspection, critical thinking, and a willingness to question (pre-existing) assumptions or beliefs. Reflective engagement often leads to a deeper understanding of concepts, thorough scrutinizing of arguments, and the ability to form well-founded opinions. In the context of argumentative ECAs, we focus on the respective argument exploration behavior of users and the possibility that this behavior is subject to a certain bias, known as confirmation bias, resulting in users trapped in their self-imposed filter bubble (SFB) [5,14,15].

Subsequently, we concentrate on conversational engagement, which refers to the level of involvement and interest exhibited by participants in a conversation. It encompasses aspects such as active listening, responsiveness, empathy, and the ability to maintain meaningful exchanges. In the context of conversational agents or dialogue systems, conversational engagement often pertains to the system's capacity to create and maintain engaging interactions with users, fostering a sense of connection and rapport. In this paper, we especially focus on the non-verbal communication behavior of ECAs, such as co-speech gestures.

2.1. Reflective Engagement in the Interaction with ECAs

Critical thinking is defined as a "metacognitive process—consisting of a number of skills and dispositions—that, through purposeful, self-regulatory reflective judgment, increases the chances of producing a logical solution to a problem or a valid conclusion to an argument Dwyer et al. [22], Dwyer [23]. According to Lucas [24] the terms critical reflection, reflective practice, reflective critical thinking and reflexivity have similar meanings and applications in educational literature [25], and are used interchangeably. Whitaker and Reimer [26] claim that the reflective process is engaged in for a purpose and within a context (e.g., social, political or ethical), where the "critical' dimension might refer to critical thinking, i.e., the weighing up of the pros and cons. Thus, critical reflection is a multifaceted and complex phenomenon involving deep knowledge and engagement in reflection processes to deepen self-awareness, better understand interaction with others, and rethink theoretical claims [27]. For example, Makhene [28] aim to facilitate engagement in argumentation and address the question of how argumentation can be used as a methodology to facilitate critical thinking in the context of nursing practitioners.

However, particularly in the field of argumentation, there is a phenomenon observable that counteracts this reflective engagement. This so-called confirmation bias [3] describes the users' seeking or interpreting of evidence in ways that are partial to their existing beliefs, expectations, or a hypothesis at hand. For example, the political polarization of social networks is the fault of "sorting based beliefs" when "viewers watch news programs and channels whose positions match their tastes and beliefs" [7].

Allahverdyan and Galstyan [29] define confirmation bias as the "tendency to acquire or evaluate new information in a manner consistent with one's existing beliefs". Furthermore, Jones and Sugden [30] demonstrated a positive confirmation bias in both information acquisition and use, as evidenced in an experiment where individuals selected "information on what to buy, prior to making a decision". Neurologically, Kappes et al. [31] illustrate an implication of confirmation bias, showing that existing judgments modify the neural representation of information strength, thereby making individuals less likely to change their opinions when faced with disagreement.

To address a user's confirmation bias within decision-making processes, Huang et al. [32] propose the use of computer-mediated counter-arguments. Additionally, Schwind and Buder [33] consider preference-inconsistent recommendations as a viable strategy to stimulate critical thinking. However, introducing too many counter-arguments can cause adverse effects such as negative emotional consequences (annoyance, confusion) [32]. According to Paul [34], when users engage in critical thinking in a "weak sense", they reflect on positions that differ from their own [35], yet tend to defend their own views without reflection [34].

Critical thinking in a "strong sense" involves reflecting on one's own opinions as well. However, the energy and effort [36] required for this strong critical reflection is often lacking due to the limited "need for cognition" [37] of people. Given users' inclination to defend their own views [34], a system that presents opposing perspectives may not necessarily foster critical reflection, but rather the opposite effect. Consequently, Huang et al. [32] emphasize the necessity for an intelligent system capable of adjusting the frequency, timing, and selection of counter-arguments. Developing such a system requires the construction of a model that can be tailored to individual users.

An approach to developing such a model is exemplified by Del Vicario et al. [38], who investigate online social debates and endeavor to mathematically model and elucidate the associated dynamics of polarization based on confirmation bias. Instead, we are focusing on modeling the underlying cause of this bias, known as the "Self-imposed Filter Bubble" (SFB) [4]. We explain our approach building upon our previous work [5,14,16] which defines measurable dimensions for describing and constructing a model of this SFB within the context of cooperative argumentative dialogue. This cooperative framework is inspired by the findings of Villarroel et al. [39], who assert that a consensual dialogue is significantly more adept at resolving conflicting perspectives on evidence and correcting incorrect, partial, and subjective interpretations of evidence compared to a persuasive dialogue. Thus, we chose a cooperative approach to exchange arguments such that the system in which our SFB Model and the respective SFB-breaking strategy are incorporated, does not try to persuade or win a debate against a user, unlike most approaches to human–machine argumentation. Those approaches utilize different models to structure the interaction and are embedded in a competitive scenario [40–46].

Instead of a persuasive approach from the previously mentioned works, our system [47] aims to engage in a cooperative exploration of arguments and to support and enhance the user's reflective engagement by building a well-founded unbiased opinion.

2.2. Conversational Engagement in the Interaction with ECAs

To transfer the inherent richness of human–human interaction to human–computer interaction and thus enhance conversational engagement, ECAs are a powerful user interface paradigm [48]. ECAs are virtual embodied representations of humans communicating multimodally with the user (or other agents) through voice, facial expression, gaze, gesture, and

body movement. A lot of prior work [49] on non-verbal communication behavior of ECAs focuses on co-speech gestures and analyses their impact on human–agent communication. McNeill's typology [50] is widely recognized and is frequently used in the classification of gestures, particularly in investigations of the cognitive aspects of gestural communication. According to this typology, gestures are divided into four primary categories: (1) pointing (deictic) gestures, (2) iconic gestures, (3) metaphoric gestures, and (4) beat gestures.

Several articles in the related literature focus on how to generate the respective ECA's behavior in a natural way [1,51–53]. In [54] a systematic review of co-speech gesture generation and evaluation methods is presented. They focus on ECAs with a human-like upper body that uses co-speech gesturing in social human–agent interaction. With advances in artificial intelligence, the methods used by these systems have evolved over the years. In this context, the question of how human gesture characteristics and theoretical frameworks on metaphors and embodied cognition can be used is analyzed by Ravenet et al. [55]. They aim to capture and understand the specific characteristics of communicative gestures in order to envision automatically generated communicative gestures. Another automatic system that parses raw text in real-time and generates appropriate emotional and gestural performance, which is claimed to also convey personality traits, is introduced by [56]. Furthermore [57] propose a framework that views text-to-gesture as machine translation, where gestures are words in another (non-verbal) language.

Binder [58] investigates the effects of body movement on conversational effectiveness in computer-mediated communication based on theories of motor cognition and embodiment. According to their definition [58,59] conversational engagement describes "the level of psychological involvement and a general state of focused activation during conversation". It can be interpreted as "the perceived commitment to and involvement in the conversation and is an indicator of the affective-cognitive resources invested in the interaction". This is undermined by the findings of Olafsson et al. [60] who showed that the interaction with the humorous agent with hand gestures, head nods, etc., led to a significantly greater change in motivation to engage in healthy behavior (increase in fruit and vegetable consumption) than interacting with the non-humorous agent.

Still, when modeling multimodal ECA behavior, the respective impact and influence on the user impression is subjective and depends on various factors. Neff et al. [61] conducted an experiment with a virtual agent that demonstrates how language generation, gesture rate and a set of movement performance parameters can be varied to increase or decrease perceived extroversion. In particular, the expressivity gesture of virtual agents has been investigated by Pelachaud [62]. Their findings suggest that co-speech gestures as one component of the expressivity of ECAs can have a significant impact on user perception, prompting us to investigate whether this is also perceivable in the context of argumentative discussions.

Therefore, we are particularly interested in whether automatically pre-defined co-speech gestures of the agent, without modeling the personality of the avatar, already imply a changed perception in the user. This study was also motivated by [63] who explored the impact of divergent interaction strategies used by a virtual recruiter on interviewees. Two distinct social behavior profiles were developed: one character adopted an understanding strategy in engaging with interviewees during the job interview simulation, while the other exhibited a markedly demanding behavior. The non-verbal behavior of the understanding agent was delineated among others by exhibiting restrained gestures proximal to the body, whereas the demanding one showed gestures that occupy more physical space (dominant gestures). Moreover, several studies [64,65] indicate that co-speech gestures have a positive impact on the learning process and user engagement in educational settings. Moreover, the results of Gratch et al. [66] in a listening condition indicate that non-verbal communication can create rapport and improve the effectiveness of a virtual agent.

In He et al. [67] the contrast between gestures produced by a machine-learning model and the idle state is examined with regard to how human participants perceived a virtual robot presenting paintings of six classical Roman monuments. Similar to our research,

they gauged user perceptions using self-assessment questionnaires, focusing on human likeness, animacy, perceived intelligence, and focused attention. Although most differences between the gesture and idle condition were negligible, the eye gaze tracker suggested that the data-driven generated gestures tended to captivate participants' attention more. These results suggest that it might be possible that users react more strongly to corresponding co-speech gestures in a direct interaction where they are actively involved. Therefore, in our study, we focused on a human-like agent engaging in direct, live conversation with the user, actively participating in turn-taking by providing spoken responses. Thereby, we were mainly interested in the user's overall perception, trust, engagement, and perceived conveyed content, as well as the impact on the user's Self-imposed Filter Bubble (SFB) and interest. As the literature on the influence of agents and their non-verbal behavior in argumentative dialogue systems remains scarce [19], existing findings lack an analysis of the change in engagement, motivation, interest, and motivation to break one's own SFB when a cooperative argumentative dialogue system features a virtual human-like agent using co-speech gestures compared to idle ones. In related work, there has not been an analysis to determine whether the behavior of co-speech gestures by a virtual discussion counterpart can influence content-related policies in argumentative discussions. This paper aims to bridge this gap by analyzing whether co-speech gestures can maintain the user's motivation during interaction, especially if the system tries to "break" the user's self-imposed filter bubble (SFB).

3. Materials and Methods

In this section, we provide an overview of the relevant models and the argument selection strategies based upon them. We first address the necessary requirements in Section 3.1, which both models require. Afterward, we explain the Interest Model introduced in [20] in Section 3.2 and the respective Interest selection strategy in Section 3.3. Following that, we explain the SFB Model, which was introduced in our previous work [5,14,16], and is described in Sections 3.4.1–3.4.3. These sections serve as a basis for our system's SFB-breaking policy [16] in Section 3.4.4.

3.1. Required Argument Structure and Clustering

In order to be able to integrate our models into our ADS and combine it with existing argument mining approaches, we adhere to the bipolar argument annotation scheme introduced by Stab et al. [68]. Due to the generality of the annotation scheme, the system is not confined to the data considered herein. In general, any argument structure that aligns with the applied scheme can be utilized. This ensures flexibility in view of the discussed topics.

This scheme 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* on the other hand are assertions that formulate a certain opinion targeting the *Major Claim* but still need to be justified by further arguments, *Premises*, respectively. Two relations between these argument components (nodes) are considered: *support* (green solid arrows) or *attack* (red dashed 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" (e.g., Claim 1) is supported or attacked by its "children". If no children exist, the node is a leaf (e.g., Premises 4–6) and marks the end of a branch.

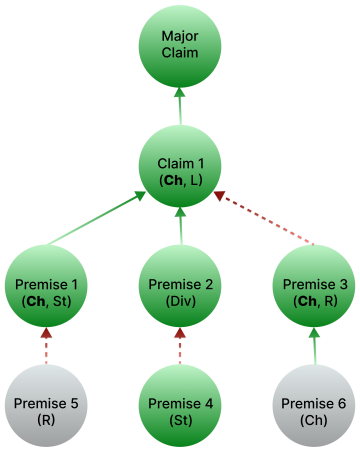


Figure 1. Visualization of exemplary argument tree structure with different argument components (visited nodes marked in green, unvisited ones in grey). The topic of the discussion is the root node, which is supported by Claim 1 (green solid arrow). Claim 1 is supported by Premises 1 and 2 and attacked by Premise 3 (red dashed arrow). The respective leaf nodes are Premise 4 (attacking Premise 1), Premise 5 (attacking Premise 2) and Premise 6 (supporting Premise 6). The according associated clusters are denoted in brackets. *Ch(ildren)* is marked in bold due to an example calculation given in Section 3.2.

Furthermore, both our Interest- and SFB Models necessitate semantically clustered arguments, such that each argument belongs to one or more content-related meta-aspects, termed "clusters", of the discussed topic. Given that an argument can encompass multiple aspects of a topic, it may belong to several overlapping clusters [69]. Thus, each argument of the argumentation tree directly addresses one or more clusters. Since each argument component targets the preceding parent, it indirectly refers to all preceding parents. Consequently, each argument component inherits the clusters of its preceding nodes, meaning that it indirectly encompasses all clusters that its parent addresses, whether directly or indirectly. Notably, the Major Claim is not affiliated with a cluster. Figure 1 exemplifies the direct cluster affiliations of the components of the argument, where "Ch" stands for children, "L" for law, "Div" for divorce, "R" for religion and "St" for relationship stability.

3.2. Interest-Model

The interest of an individual user can be determined through explicit or implicit user feedback. Typically, systems that collect implicit data impose minimal or no burden on the user, making them more likely to be adopted [70]. Thus, our aim is to implicitly identify: (1) whether the user has lost interest I_{loss} in the currently discussed branch, and (2) which claim would best suit the user's interest to be presented after an interest loss I_{loss} is detected.

Most approaches to estimating user interest typically rely on explicit user feedback, but we opt for an implicit approach to ensure a natural, content-driven dialogue. Existing implicit methods focus on modeling user interest in website content, using factors like browsing history and reading time, with little consideration for dialogue systems. Drawing parallels, our user Interest Model is based on insights from the work of Yi et al. [71]. They suggest that shorter website content and longer browsing times indicate greater user interest. Similarly, our model posits that longer conversations imply a stronger interest in the topic discussed. Importantly, "length" here does not refer to temporal duration, as the responses of the ADS, i.e., the presented argument components, vary in length, which the user cannot predict in advance. Additionally, some available user moves (e.g., general interaction information) are not content-related and should be excluded from the Interest Model.

Therefore, we focus on the ratio of requested arguments in a specific subtopic cluster to the total available arguments. In line with the approach presented by [71], we consider content-specific branches of our argument tree structure. As detailed in Section 3.1, each

node inherits the clusters of its predecessors due to the logical structure where each argument component builds on the previous. Hence, the length of the visited argument branch is an important factor.

Every argument (except for "leaf nodes") directly addressing a cluster serves as the root node of a relevant subtree considered in the calculation. Subtrees assigned to the same cluster can overlap, counting multiple times, especially when an argument component directly addresses a cluster that is already addressed by one of its predecessors (see Figure 1 for the cluster *Children*). Thus, instances where the user explicitly requests further information on an already introduced cluster are considered as they indicate increased interest in that topic.

In the following sections, we explain our overall Interest Model and the argument selection strategy based on it (the "Interest strategy"). For a detailed explanation of its components, we refer to the Appendix A.1.

To determine the overall user interest in a specific cluster k, all subtrees of the argument structure tree that belong to k must be considered. Therefore, we iterate over all root nodes of the subtree r_k and aggregate all weighted subtree interest values. The overall user interest for cluster k is given by

$$I_k = \frac{\sum_{r_k} \omega_{d,B_{r_k}} \omega_{d,B_{r_k}} I\left(B_{i_k}\right)}{\sum_{r_k} \omega_{d,B_{r_k}} \omega_{d,B_{r_k}}}. \quad (1)$$

The interest values in Equation (1) are normalized so that $I_k \in [0,1]$. Hence, $I_c = 1$ denotes the highest possible user interest in cluster k, whereas $I_k = 0$ denotes the lowest.

For a better understanding, an example calculation with respect to the argument tree shown in Figure 1 is provided in the Appendix A.2.

3.3. Interest Selection Strategy

This strategy consists of two components. First, using the Interest Model described in Section 3.2, it should be determined whether the user has lost interest (I_{loss}) in the current branch. Then, it should be determined which claim interests the user the most and should be presented to them.

As detailed in [20] in order to recognize I_{loss}, we trained a binary neural network classifier to predict whether users remain interested in the content of the ongoing dialogue. As input features, the calculated user interest values of our Interest Model as well as the number of visited arguments per cluster were used. This resulted in a classification accuracy of 74.9%.

Moreover, we investigated how the system can predict which claim might be best to present to the user after I_{loss} is detected. Various approaches on ANN multi-class classification were explored, but none of them yielded reliable results. Therefore, we developed a rule-based approach, leveraging our user Interest Model detailed in Section 3.2, to select an appropriate next claim.

The procedure is as follows. All unvisited claims are considered potential candidates. If only one claim remains unvisited, it is presented to the user. Otherwise, we prioritize the claims of the cluster with the highest interest value. However, it is possible that more than one unvisited claim belongs to the same cluster. In such cases, we select a claim that belongs to clusters with high user interest and avoids uninteresting clusters. To achieve this, we calculate the average interest for each eligible claim across all the clusters it belongs to and then choose the claim with the highest value. If there are still multiple claims remaining, we need another selection criterion.

Thus, the user's preference for specific arguments and their cluster association is taken into account. Each time an argument is preferred by the user (see Section 1), the counters $pref_k$ for the associated clusters k increase. Similarly, each time an argument is rejected by the user, the counters rej_k for the respective clusters k increase. Subsequently, the preference quotient λ_k can be calculated:

$$\lambda_k = \frac{pref_k - rej_k}{n_k}, \tag{2}$$

where n_k denotes the total number of arguments belonging to k. Therefore, the claim is selected with the summed preference quotient:

$$\lambda_{claim} = \sum_{k \in K} \lambda_k, \tag{3}$$

it is the highest, where K is the set of cluster indices to which the claim belongs. If there are still multiple claims remaining, a random one is chosen and presented to the user.

3.4. Self-Imposed Filter Bubble

In this subsection, we provide an overview of the SFB Model we adapted after explaining its corresponding dimensions in Section 3.4.1. This serves as the foundation for our system's SFB-breaking policy introduced in Section 3.4.4.

3.4.1. SFB Model Dimensions

We adapted the SFB Model introduced by Aicher et al. [5] which is motivated by the "Elaboration Likelihood Model" (ELM) [72]. As already mentioned, it incorporates four dimensions, which span a four-dimensional space to describe the user's SFB, which are described in detail in the Appendix A.3: *Reflective User Engagement (RUE)*, *Personal Relevance (PR)*, *True Knowledge (TK)* and *False Knowledge (FK)*. We explain the basic model of how these dimensions form the clusterwise and overall SFB vector of the user and necessary requirements for integration in an ADS. Please note that we do not claim that the dimensions or our model are complete, but it is a first approach to model SFBs.

Argumentative discussions are intricate, encompassing diverse subtopics containing arguments pertaining to similar content-related aspects (clusters). For each of these clusters, we define corresponding SFB-vectors $\overrightarrow{sfb_k}$, $k \in \mathbb{N}$ (one for each subtopic), which collectively constitute the overall SFB-vector $\overrightarrow{SFB_k}$ of the entire discussion topic.

It is imperative to differentiate between the SFB and SFB-vector of a user (see Figure 2). The SFB vector is conceptualized as a vector originating from the origin of the coordinate system and ends at the user's position in the four-dimensional space at the current interaction state. Additionally, the SFB encompasses areas in the four-dimensional space indicating the probability that users are caught within an SFB when their SFB-vectors lie in this area depending on predefined limits.

3.4.2. Clusterwise SFB Model

With the previously defined dimensions and derived Equations (A7)–(A10) and (A11), the user's SFB-vector for each cluster k is obtained:

$$\overrightarrow{sfb_k} = (pr_k, rue_k, tk_k, fk_k)^T. \tag{4}$$

In Figure 2, an exemplary sketch of two vectors and the respective SFB is shown. As a four-dimensional space cannot be displayed, it has been transformed into a three-dimensional space by merging the dimensions of True Knowledge and False Knowledge into the Knowledge (K) dimension. Please note that this sketch is for illustrative purposes only, and the "real" shape and structure of the blue SFB sphere may differ. Especially, as it is hard to define distinct margins, we describe a probability for a user to be inside or outside the SFB.

Depending on the definition of "breaking" the SFB for a cluster, using this vector, various criteria could be examined. In the first step, we compare the initial (before the interaction) and final (after the interaction) SFB-vector position concerning the clusterwise SFB. For instance, a minimum relative change $\delta_{k,min}$ between the initial and final position could be defined and compared by calculating the difference in magnitudes δ_k:

$$\delta_k = |\overrightarrow{sfb_{k,f}}| - |\overrightarrow{sfb_{k,i}}|, \tag{5}$$

where $\overrightarrow{sfb_{k,i}}$ consists of the initial values (especially $tk_{k,i}$ and $fk_{k,i}$ have to be estimated, e.g., initialization dialogue or questionnaire, before and verified during the interaction). $\overrightarrow{sfb_{k,f}}$ denotes the final vector after the interaction ended. If $\delta_k > \delta_{k,min}$, there is a high probability that the users find themselves outside the clusterwise SFB. However, since these considerations only include one cluster, it is necessary to consider all clusters.

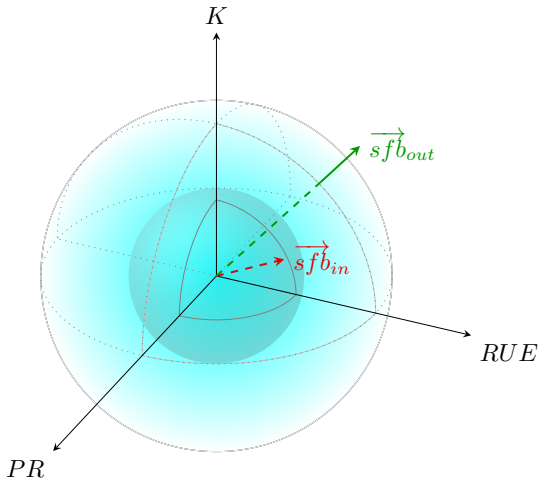


Figure 2. Schematic sketch of the SFB and two SFB-vectors, one inside (red dashed arrow) $\overrightarrow{sfb_{in}}$ and one outside (green dashed and solid arrow) $\overrightarrow{sfb_{out}}$ of the SFB. The blue sphere indicates the SFB of a user, representing the probability that the user is in an SFB. The closer the SFB-vector of the user ends to the origin (when the SFB dimensions are close to zero), the higher their probability of being caught in a filter bubble, and vice versa. This is illustrated by the color gradient from dark and intense to light and transparent blue. Thus, this probability for $\overrightarrow{sfb_{in}}$ is quite high and for $\overrightarrow{sfb_{out}}$ quite low. The axes represent the three dimensions: Personal Relevance (PR), Reflective User Engagement (RUE), and Knowledge (K). For illustration purposes, the four-dimensional space has been transformed into a three-dimensional space by merging the dimensions True Knowledge (TK) and False Knowledge (FK) into the Knowledge (K) dimension. All values are normalized.

3.4.3. Overall SFB Model

In order to define an overall SFB Model for all clusters, the differences between these clusters have to be taken into account regarding the determination of the Reflective User Engagement. When considering hierarchical argumentation structures, e.g., argument trees, arguments at the beginning of a branch are more general than ones at deeper levels. Due to this, we introduce a hierarchical weight ω_d to incorporate the different levels of argument depth into the overall RUE measure. Therefore, a balanced exploration of lower argumentation levels will be assigned larger weights than near the root node (see Figure 1). As the depth of arguments within the argument tree may vary, we define a median depth $d_k = \text{med}(D)$ with D denoting all depths of the respective visited arguments belonging to cluster k (taking the average instead would lead to a great bias, especially as $d \in \mathbb{N} \quad \forall\, d \in D$). Thus it follows:

$$\omega_{d,k} = \frac{d_k}{\sum_{m=1}^{d_{k,\max}} m}, \tag{6}$$

with $d_{k,\max} = \max(\text{med}(d_k)), k = 1 \ldots n$ being the maximum median depth of all n clusters.

Furthermore, to avoid an over-representation of clusters with only a few arguments while clusters with many arguments will be under-represented, we define a weight $\omega_{k,n}$

which takes the different sizes of clusters into account. Thus, the number of arguments n_k within the cluster k is related to all arguments in all clusters n_{all} such that:

$$\omega_{n,k} = \frac{n_k}{n_{\text{all}}}. \tag{7}$$

By merging Equations (A7), (6) and (7) and respective normalization, the overall RUE for all n clusters can be determined by:

$$RUE = \frac{\sum_{k=1}^{n} \omega_{d,k}\omega_{n,k} rue_k}{\sum_{k=1}^{n} \omega_{d,k}\omega_{n,k}}, \tag{8}$$

with $RUE \in [0,1]$. An RUE equal to 0 indicates a strong SFB, whereas an RUE equal to 1 indicates the opposite.

Concerning the other dimensions, we take the respective average over all clusters, such that:

$$X = \frac{\sum_{k=1}^{n} x_k}{n}, \tag{9}$$

with $X \in \{PR, TK, FK\}$ and $x \in \{pr, tk, fk\}$.

Likewise, to the clusterwise SFB, we obtain an overall SFB-vector $\overrightarrow{SFB} = (PR, RUE, TK, FK)^T$, consisting of the overall cluster values for each dimension. This vector can serve as a starting point to determine the probability with which the user is caught within an SFB on the whole topic. Thus, we define the probability to be within an SFB as:

$$\begin{aligned} |\overrightarrow{SFB}| < \zeta_1 &: \quad \text{high} \\ \zeta_1 \leqq |\overrightarrow{SFB}| \leqq \zeta_2 &: \quad \text{moderate} \\ |\overrightarrow{SFB}| > \zeta_2 &: \quad \text{low.} \end{aligned}$$

ζ_1 and ζ_2 are margins between 0 and 1, which have to be calculated according to how strictly the SFB is defined. This is further explained in the following Section 3.4.4.

3.4.4. SFB-Breaking Policy

Building upon the previously introduced SFB Model in [16] we define a rule-based system policy with the objective of breaking the user's SFB. Utilizing data from a prior crowd-sourcing user study, we investigated how SFB dimensions changed under two distinct system policies. The first policy, as outlined in Section 3.2, follows the interest-based approach, selecting arguments based on the estimation of the user's greatest interests. The second policy involves the random presentation of arguments from the remaining set. The calculated averages across all participants were utilized as benchmark values for identifying regions where there is a higher probability of being caught in an SFB (very high probability = interest average =: ζ_1; medium probability = random average =: ζ_2).

Given that PR and FK cannot be ascertained beforehand but only in hindsight, the rule-based policy focuses on maximizing the RUE and TK dimensions, which can be computed in advance. If the values for PR or FK deteriorate (become smaller) after introducing a new argument, we assign a greater weight to the associated cluster and respective arguments to counteract this.

To ensure logical coherence, it is important that potential argument candidates are logically connected to the requested argument, either through sibling relationships or by sharing the highest degree of overlap in their respective cluster affiliations. Once candidates are identified, they are evaluated against the user-selected argument in terms of the corresponding RUE and TK dimensions. Subsequently, the argument with the maximum values in these dimensions is presented. In cases where the system selects an argument different from the user's choice, the system response includes an explanation so that the user understands the system's choice.

Following an initialization phase (first five argument requests) aimed at detecting and rewarding shifts in users' exploration behaviors, the user's current SFB vector is compared to the data-based SFB margins (interest, random) after each interaction turn. If the SFB vector falls within the first area (below the interest margin), the ADS will consistently opt to select the best available argument in each turn. When the SFB vector is situated within the second region (above the interest margin, below the random margin), a decision is made based on recent changes in the SFB vector over the preceding three interaction turns. This determines whether the system offers an "SFB-breaking" argument or the requested argument. If the SFB-vector surpasses the random margin, the ADS presents the requested argument, contingent upon the precondition that the absolute value of the SFB-vector did not decrease in the preceding turn.

4. ADS Architecture

In the following, the architecture of our argumentative dialogue system and its components, in particular explaining the further requirements that need to be fulfilled to incorporate our previously described models in Sections 3.2 and 3.4 into the underlying dialogue model. Furthermore, the interface and non-verbal behavior modeling of the agent are outlined.

4.1. Dialogue Model

As pointed out in Section 3.1 the arguments are required to fit a bipolar argument tree structure following the annotation scheme of Stab et al. [68] and need to be clustered. These conditions are fulfilled by the sampling debate on the topic *Marriage is an outdated institution* that is used in the discussed study. It serves as a knowledge base for the arguments and is taken from the *Debatabase* of the https://idebate.org/debatabase (accessed on 23 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.) 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 [73] for further use. In this dataset the maximal depth of a branch d_{max,B_j} varies from 5 to 10. With regard to the argument clustering, the following ten clusters were identified in our sample dataset: *Alternative relationships and parenthoods*, *Children*, *Divorce*, *Expectations and commitment*, *Harmful relationships*, *Law*, *Relationship stability*, *Religion*, *Remarriage, and Social Acceptance*. In each *why pro/con* move, a single supporting/attacking argument component is presented to the user according to the respective argument selection strategy. To prevent the user from being overwhelmed by the amount of information, the available arguments are presented to the users incrementally upon their request.

In order to integrate the Interest Model in Section 3.2 and the SFB Model in Section 3.4, the dialogue model has to provide respective user moves. The interaction between the system and the user is separated into turns, consisting of a user action and the corresponding natural language answer from the system. The system's response is based on the original textual representation of the argument components, which is embedded in moderating utterances.

Table 1 shows the required (only moves that are relevant for the Interest and SFB Model are shown. Other moves are not listed due to their mere navigational/meta-informational purposes.) possible moves (actions) the user is able to choose from. This allows the user to navigate through the argument tree and inquire for more information. The determiners indicate which moves are available depending on the position of the current argument (root node, parent node, or leaf node).

Table 1 shows which user moves affect which components of the respective models. With the exception of *level up*, which only affects the Interest Model components, the remaining moves affect both. This holds especially for I_{loss} as it is trained with a feature set containing the previous user actions/moves. Whereas why_{pro}, why_{con} and *suggest* in-

dicate a low interest loss I_{loss}, $prefer$, $reject$, $know$, $false$ and $level\ up$ indicate the opposite. Furthermore, regarding the SFB dimensions, especially, rue_k, t_k, and f_k are directly influenced by respective user moves and thus updated immediately. However, this does not apply to PR, which does not directly refer to the dialogue content but rather serves as a meta-reflection. Since pr_k does not directly pertain to the argument, but rather to the respective cluster, this information is requested in a separate pop-up window during the interaction as shown in Figure 3. To avoid inconveniencing the user (given that the cluster might remain the same over a certain number of moves), we update pr_k whenever the corresponding clusters change (when a new cluster k_{new} is addressed and the old cluster k_{old} is no longer addressed).

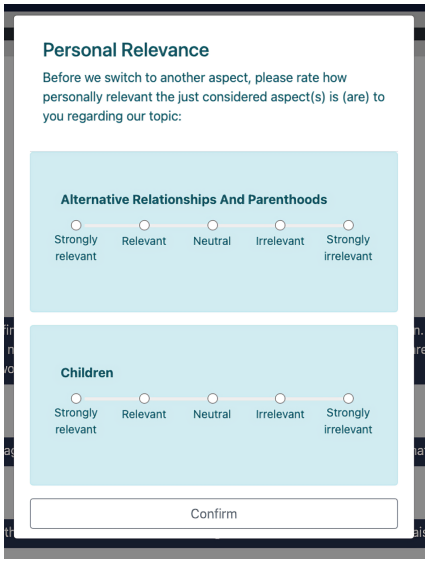


Figure 3. Exemplary Popup for the PR 5-point Likert rating of the clusters "Alternative relationships and parenthoods" and "Children" which were previously addressed before the switch to another cluster.

Table 1. Description of the possible user moves with corresponding determiners and influenced SFB dimensions and Interest Model components. The latter is updated dynamically after each move.

Move	Description	Determiners	Interest Comp.	SFB Dim
why_{pro}	Requesting for a pro argument	If supporting child exists	I_{loss}, $\lvert B_{i_k,v}\rvert$	rue_k, tk_k
why_{con}	Requesting for a con argument	If attacking child exists	I_{loss}, $\lvert B_{i_k,v}\rvert$	rue_k, tk_k
$suggest$	Requesting another argument	If unheard arguments exist	I_{loss}, $\lvert B_{i_k,v}\rvert$	rue_k, tk_k
$prefer$	Agree/prefer argument	Always	I_{loss}, λ_k	rue_k
$reject$	Disagree/reject argument	Always	I_{loss}, λ_k	rue_k
$know$	Stating argument is already known	Always	I_{loss}	$tk_{k,i}$ [1]
$false$	Stating argument is incorrect	Always	I_{loss}	fk_k
$level\ up$	Request to switch to "parent"	If parent exists	I_{loss}	

[1] The index i refers to the initial true knowledge (before the interaction).

An exemplary dialogue for each argument selection strategy is shown in the Appendix B.

4.2. Interface and Non-Verbal Agent Behavior

The interface is based and extended from our argumentative dialogue systems (ADS) introduced in previous work [21]. The interface for all four study conditions (see Section 5), depicted in Figure 4, is completely identical, except for two differences. The first one is the two different argument selection strategies (Interest, SFB-breaking), and the second one

is the non-verbal behavior of the agent. Examples of a gesticulating agent are illustrated in Figure 5. The interface is based on the Charamel^{TM} avatar (licensed under CC BY 4.0, https://creativecommons.org/licenses/by/4.0, accessed on 23 November 2023) which presents the system utterance by lip-sync speech output using the Nuance TTS along with the Amazon Polly voices (https://docs.aws.amazon.com/polly/latest/dg/voicelist.html, accessed on 23 November 2023).

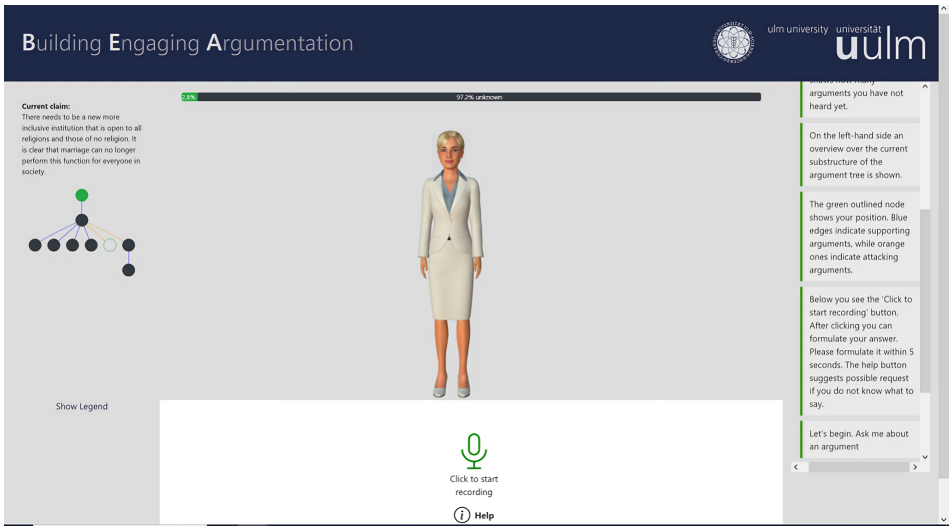


Figure 4. GUI of the no-gesture systems. Above the button "Click to start recording" the agent is shown. The dialogue history is displayed on the right side of the screen, system utterances are marked in green and user responses in blue. On the left side, the sub-graph of the current branch is visible. Above the agent, a progress bar is shown.

Figure 5. Exemplary co-speech gestures of the agent in the gesture systems. The image on the (**left**) and in the (**middle**) point to a GUI element referenced in the interaction. The image on the (**right**) depicts one of the 25 predefined explanatory speech gestures.

We opted for a full-body representation of the agent (in the middle of the GUI) as it moves across the screen to introduce and highlight various elements of the GUI. This furthermore enabled us to make use of the more than 50 conversational motion-captured gestures supplied by `Vuppetmaster` (https://www.charamel.com/products/vuppetmaster, accessed on 10 June 2024) which can be adjusted during run-time in some aspects (e.g., overall speed, extension, etc.). As shown by Pelachaud [62] modifying the "spatial extent (quantity of space taken by a body part)" and "temporal extent (velocity of execution of a movement)" are noticed most by users and influence the expressivity and thus perception of personality of the agent. Aiming to maintain a neutral impression of the agent, we do not

manipulate these "expressivity dimensions" [62] but rather utilize the settings provided by the `Vuppetmaster`. However, this aspect should be further investigated in future work. Our choice is further motivated by the work of [63] as they showed that the behavior of a virtual character in a recruiting scenario did have an impact on the participants' subjective experience and their performance in the simulation.

In addition, the character comes with a catalog of 14 facial expressions, which contains among others the six basic emotion expressions defined by Ekman [74]. It is important to note that the focus of the study was to examine the influence of an agent using 'suitable' co-speech gestures (movements of arms and hands for explanation, head movements, etc.), which emphasized the verbal introduction to the arguments and their presentation. The study did not include an analysis of the influences of explicit facial expressions/emotions. Due to this, and the fact that the agent is intended to be perceived as a neutral and unbiased conversational partner, we restricted ourselves to neutral facial expressions. To ensure the suitability of the co-speech gestures for our purpose, they were manually selected from the set of available conversational motion-captured gestures. In this process, we adhered to criteria defined by two independent experts as "natural and appropriate for an argumentative discussion with a neutral conversational partner". These criteria are as follows:

- No large leg movements (*jumping*, *hopping*, *dancing*, etc.); lateral steps are allowed.
- No turning of the upper body and face away from the user at an angle greater than 45 degrees.
- Movements of the torso are allowed as long as they are not fast, hectic, jerky, or incompatible with the flow of conversation.
- Hand and arm movements are limited to the area of the torso, not above shoulder height unless explicitly pointing to an object above.
- No movements that can be interpreted in the context of emotions (e.g., stomping the foot or waving) or indicate a non-neutral conversational partner (e.g., crossing arms, thumbs up).

The co-speech gestures determined according to these selection criteria were not customized to the specific content of the arguments or adapted to individual users. Instead, they were adapted to the content of the agent's statement in the context of the dialogue but remained identical for each user if the corresponding statement was presented. More precisely, for each user move (see Table 1), corresponding co-speech gestures were defined, which are aligned with the moderating agent's utterance. Moderating agent utterances serve to embed the response within the context of the interaction and to pick up on what the user said before the agent provides the requested information. The selection was manually assigned to ensure high relevance, coherence, and consistency. For instance, if a new argument was requested, one of the 25 co-speech gestures predefined by the `Vuppetmaster` was randomly chosen. This random selection ensures that no gesture is associated with a specific polarity or cluster. Furthermore, we took care that the same speech gesture was not used in the previous five turns. The synchronization of co-speech gestures with the utterance was also handled by the `Vuppetmaster`. Only one speech gesture was selected for each agent turn to avoid overloading the interaction. The neutral representation of the argument is supported by an explanatory, non-polarizing co-speech gesture of the avatar (see Figure 5 on the right). By selecting from 25 co-speech gestures, it is ensured that the agent's gestures are not repetitive and thus appear natural. However, if the user move clearly refers to an element found in the GUI, this will be emphasized in the corresponding co-speech gesture (see Figure 5 left, middle). An example of this would be the user's statement to revisit a previously presented argument, which the agent indicates by pointing to the dialogue history.

The dialogue history is shown on the right side of the screen, marking the system answers with a green and the user answers with a blue line. 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 green line. Already heard arguments are shown in blue. Nodes shown in grey are still unheard of. 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.

An NLU framework based upon the one introduced in previous work [75] processes the spoken user utterance. By clicking on "Click to start recording" the user starts the recording and can formulate their request within 5 s after which the recording automatically stops. The spoken input is captured by a browser-based audio recording that is further processed by the `Python` library `SpeechRecognition` (https://pypi.org/project/SpeechRecognition/, accessed on 17 July 2023) using the Google Speech Recognition API. Its intent classifier uses the BERT Transformer Encoder presented by Devlin et al. [76] and a bidirectional LSTM classifier. The system-specific intents are trained with a set of sample utterances from 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.

5. User Study Setting

We conducted a user study with 56 participants (aged 22–41; 15 female, 41 male) divided into four groups:

- Group I+nG (15 participants):
  Argument selected according to Interest Model (SFB); no co-speech gestures (nG);
- Group SFB+nG (14 participants):
  Argument selected according to SFB Model (SFB); no co-speech gestures (nG);
- Group I+G (13 participants):
  Argument selected according to Interest Model (I); co-speech gestures (G);
- Group SFB+G (14 participants):
  Argument selected according to SFB Model (SFB); co-speech gestures (G);

With regard to argumentative content, Group I+G and Group I+nG were presented with arguments based on their interests (referred to as the "I" groups). Group SFB+G and Group SFB+nG on the other hand were presented with arguments that aimed to challenge their existing beliefs (referred to as the "SFB" (breaking) groups). In the SFB-breaking (SFB) groups, the system presented arguments based on the system policy described in Section 3.4.4. Consequently, the arguments presented to the SFB-breaking group might have differed in polarity and/or cluster from the original user request. In the interest (I) groups, the system presented arguments that precisely matched the user's requests. If a loss of interest (see Section 3.2) was detected the system suggested arguments that aligned best with the user's preferences and interests. This interest policy is based on our Interest Model [20] introduced in Section 3.2. With regard to the avatar gesture animation (see Section 4.2), Group SFB+G and Group I+G were presented with a human-like avatar using co-speech gestures (referred to as the "gesture" (G) groups) whereas Group SFB+nG and Group I+nG (referred to as the "no-gesture" (nG) groups) interacted with a static avatar without any co-speech gestures.

The primary objective of the herein-discussed study is to address the following four research questions:

RQ1 How does the applied argument selection strategy influence the user's overall perception of the agent, engagement and opinion building?

RQ2 How does the presence of co-speech gestures influence the user's perception of the agent, engagement and opinion building?

To investigate these research questions, we formulated the following hypotheses to be tested in the study:

H1 The perception of the provided content is influenced positively by the co-speech gestures (especially with regard to the SFB+G group).
H2 Participants of the two SFB-breaking (interest) groups do not show a significant difference in the efficiency of breaking the user's SFB.
H3 Participants of all four groups do not show a significant difference in their opinion forming.
H4 Participants of all four groups do not show a significant difference in their self-assessed trust.
H5 Participants in the gesture and no-gesture groups show less difference within their groups (SFB+G vs. I+G, SFB+nG vs. I+nG) in terms of their engagement and general perception of the agent than compared to the groups with similar-argument-choice (SFB+G vs. SFB+nG, I+G vs. I+nG).

These hypotheses were formulated to assess the impact of the presented argument selection strategies and co-speech gestures in enhancing user engagement and motivation when interacting with the argumentative virtual agent. This is particularly relevant when the argument selection strategy is designed to challenge and break the user's filter bubble.

The study took place in a university laboratory and involved participants from diverse international backgrounds with a proficient level of English. The entire study, from the introduction to the completion of pre- and post-questionnaires, was expected to take approximately one hour. Participants received a compensation of $10 ($10 per h).

This study utilized a speech-based input and output modality, allowing participants to interact with the system through spoken language. After a brief system introduction, including instructions on interaction, participants were required to answer two control questions to ensure their understanding. Only those who passed this test proceeded to the test interaction with the system. In this test scenario, participants familiarized themselves with the system until they felt confident enough for the "real" interaction.

During the "real" interaction, participants were tasked with listening to a minimum of 20 arguments, ensuring ample data collection. Participants were unaware of the two different argument selection strategies or the presence/absence of co-speech gestures by the avatar. They were only informed that the ADS might suggest arguments on its own, and they had the option to revert to the previous argument if not approved.

In order to accurately capture the user's perception and sentiments regarding the content, presentation, and overall impression at each step of the interaction, a popup window was displayed after every turn (user move and system response). As this constitutes a meta-analysis of the ongoing dialogue, it was ensured that the intermediate feedback is presented not within the dialogue with the agent but in a popup window, as illustrated in Figure 6.

Using a numeric keypad, users could select which of the response options (*no*, *neutral*, and *yes*) applied to the last interaction turn in relation to the following three questions:

Q1 Content: Does the argument (content) contribute to the discussion?
Q2 Manner: Is the presentation of this argument (content) motivating/engaging?
Q3 General Impression: Do you feel satisfied with the dialogue at this point?

Throughout the study, the following data were collected: Self-assessment questionnaires [77–79] calculated user interest and interest loss, and calculated SFB-values: *RUE*, *PR*, *TK*, and *FK* (for each cluster k), turnwise participant ratings, participants' opinions and interests regarding the topic of discussion (via calculation and self-assessment rating), set of heard arguments and dialogue history. Furthermore, the complete interaction was captured on video using a webcam on top of the screen. Strict adherence to data protection regulations and participant anonymity were maintained throughout the study. Participants had the freedom to withdraw from the study at any time.

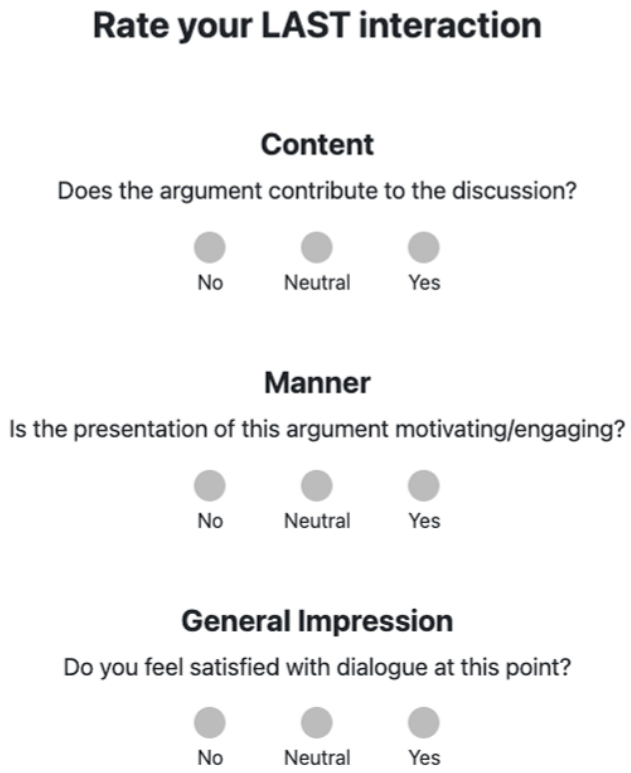


Figure 6. Popup window for intermediate user feedback on the provided content, the manner and the general impression.

6. Results

In the following section, we present the results of the previously outlined user study. In Section 6.1 the differences in the SFB dimensions for all groups are shown. Section 6.2 presents the differences in the intermediate turnwise user ratings, followed by the evaluation of the self-assessment questionnaires in Sections 6.3, 6.3.2, 6.3.3 and 6.5.

Each $\sum$ represents the summation of all items of the corresponding aspect, or if marked with (*), their respective inverted counterparts.

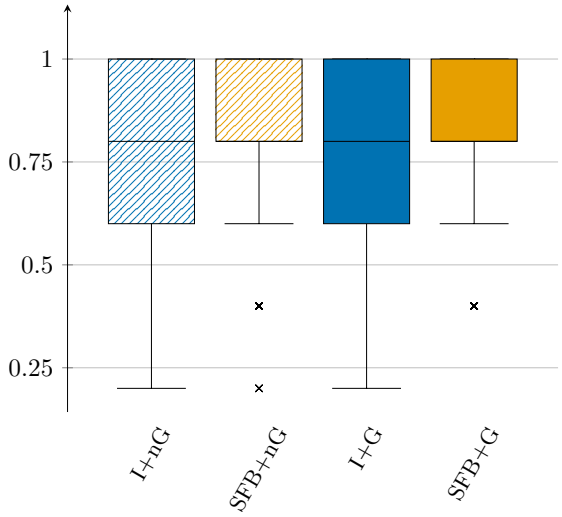
On average, participants interacted with the ADS for an approximate duration of 33 min and 41 s (SD: 6 min and 49 s) while listening to around 22 arguments.

6.1. Breaking of SFB

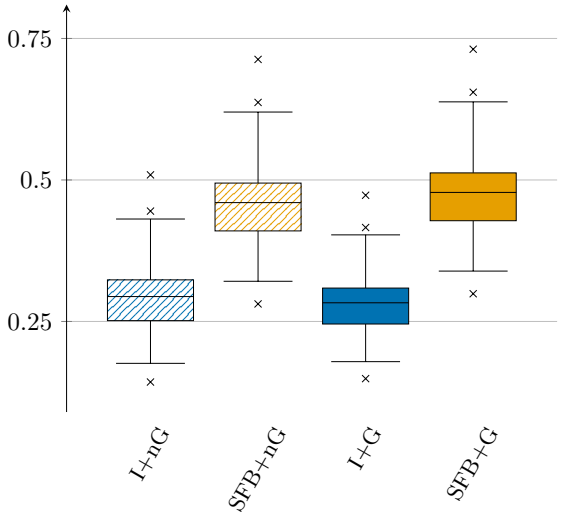
In the following, we present the results for all four SFB dimensions, the calculation of which is described in Section 3.4. As described in Section 3.4, the range of Reflective User Engagement, True and False Knowledge is continuous within the interval $(0, 1]$. However, the range of the Personal Relevance is discrete, specifically {0.2, 0.4, 0.6, 0.8, 1}. In Figure 7, we display the mean values for all dimensions across all clusters for each of the four groups. We focus our analysis on weighted overall means for each SFB dimension, calculated by averaging across all clusters since these values are representative of each individual cluster. As already evident in Figure 7, statistically significant differences ($p < 0.001$) can be observed between the SFB-breaking and interest groups using the Kruskal–Wallis test in all overall SFB dimensions among the four groups.

A Dunn–Bonferroni corrected pairwise comparison of the groups in Table 2 shows a significant difference between the SFB-breaking and interest groups for all four overall SFB dimensions. Only between the groups with the same argument selection strategy, the differences are not statistically significant. The effect is particularly strong for the dimension RUE (see also Figure 7b) with r = 0.605–0.724. For PR medium to high effect

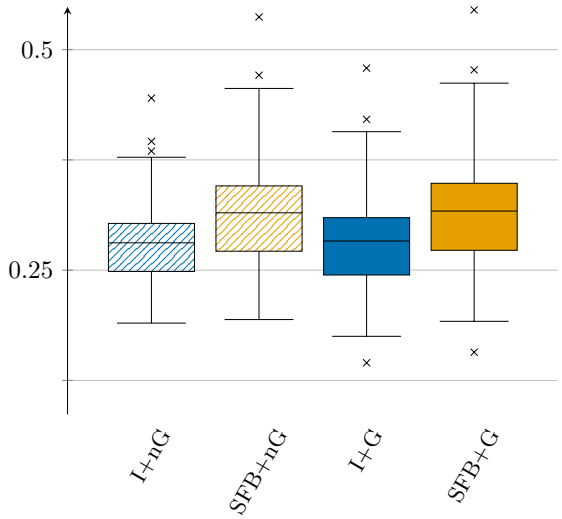
sizes are perceived (r = 0.413–0.723) as well as for TK (r = 0.323–0.580). Regarding FK the effect is small to medium with r = 0.372–0.472.



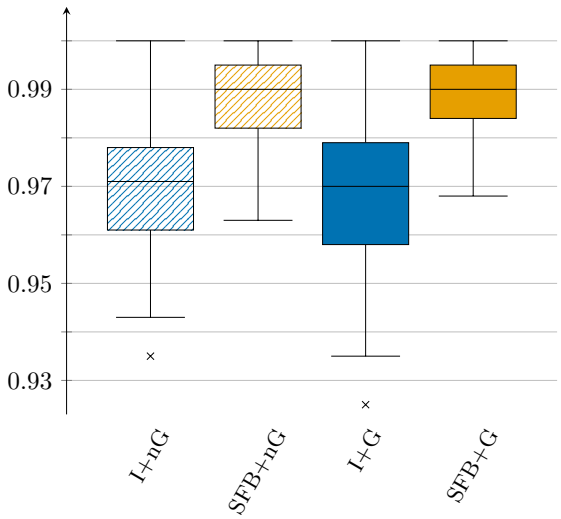
(**a**) Boxplot for overall SFB dimension *Personal Relevance*.



(**b**) Boxplot for overall SFB dimension *Reflective User Engagement*.



(**c**) Boxplot for overall SFB dimension *True Knowledge*.



(**d**) Boxplot for overall SFB dimension *False Knowledge*.

Figure 7. Subfigures (**a**–**d**) show the boxplots for each overall SFB dimension for each experimental group.

Table 2. Means and SD of all SFB dimensions over all cluster for all four groups. Significant differences ($\alpha = 0.05$, Dunn–Bonferroni corrected) in pairwise group comparisons are indicated by a bold p value.

Asp.	I+nG		SFB+nG		I+G		SFB+G		I+nG &	I+nG &	I+nG &	SFB+nG &	SFB+nG &	I+G &
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	SFB+nG	I+G	SFB+G	I+G	SFB+G	SFB+G
PR	0.776	0.21	0.803	0.18	0.778	0.19	**0.804**	0.17	**<0.001**	1.00	**<0.001**	**<0.001**	1.00	**<0.001**
RUE	0.291	0.26	0.472	0.26	0.305	0.29	**0.473**	0.26	**<0.001**	1.00	**<0.001**	**<0.001**	1.00	**<0.001**
TK	0.278	0.19	0.312	0.25	0.280	0.18	**0.315**	0.25	**0.001**	1.00	**<0.001**	**0.001**	1.00	**<0.001**
FK	0.971	0.09	0.990	0.07	0.970	0.09	**0.991**	0.05	**0.002**	1.00	**0.001**	**0.001**	1.00	**0.001**

6.2. Turnwise User Ratings

As described in Section 5 the participants rated the ongoing interaction after each turn with regard to Content (Q1), Manner (Q2) and General Impression (Q3). The corresponding data distribution for each question is depicted in Figure 8.

(**a**) Boxplot for participant ratings of Q1. (**b**) Boxplot for participant ratings of Q2. (**c**) Boxplot for participant ratings of Q3.

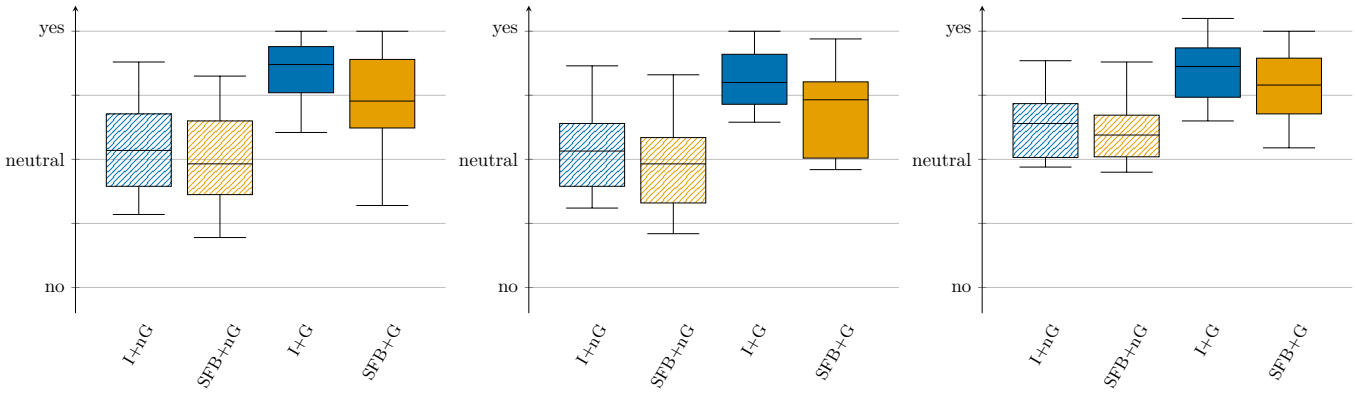


Figure 8. The three Subfigure (**a**–**c**) show the boxplot for the turnwise user ratings on the scale "no", "neutral" or "yes" regarding the three Questions Q1–Q3.

All user ratings, after the completed initialization phase. From this point, there was a difference between the SFB and Interest groups regarding the argument selection strategy. Five presented arguments were averaged for each user and tested for statistically significant differences between the four groups (I+nG, SFB+nG, I+G, SFB+G) using the Kruskal–Wallis test. In all three questions, very significant differences ($p < 0.001$) were observed, leading to the presentation of Dunn–Bonferroni corrected pairwise group comparisons in Table 3. Significant differences are indicated by a bold p-value.

As evident in Figure 8a–c, the gesture and no-gesture groups differ significantly, which is underscored by the results of pairwise comparisons. Regarding whether the argument (content) contributed to the discussion at the current point of the interaction, Table 3 shows that there is a highly significant difference with high effect sizes between SFB+nG and I+G ($r = 0.544$), as well as between SFB+SFB+G and SFB+G ($r = 0.511$). The pairwise comparisons for Q1 did not reveal any significant differences.

Table 3. Pairwise group comparisons of turnwise user-ratings of Q1–Q3. Significant differences ($\alpha = 0.05$, Dunn–Bonferroni corrected) are indicated by a bold p-value.

Question	I+nG & SFB+nG	I+nG & I+G	I+nG & SFB+G	SFB+nG & I+G	SFB+nG & SFB+G	I+G & SFB+G
Q1: Content	0.665	0.201	0.162	<0.001	**<0.001**	1.00
Q2: Manner	0.611	**0.019**	**0.023**	**<0.001**	**0.001**	1.00
Q3: General Impression	1.00	**0.009**	**0.010**	**<0.001**	**0.001**	1.00

For Q2 and Q3, only within the no-gesture (I+nG and SFB+nG) and gesture (I+G and SFB+G) groups, no significant differences were observed. Regarding the question of whether the presentation of the current argument (content) is motivating/engaging (Q2), significant differences with medium effect sizes can be perceived between I+nG and I+G ($r = 0.378$), similarly to I+nG and SFB+G ($r = 0.308$). Moreover, the differences between SFB+nG and I+G, as well as SFB+nG and SFB+G, are highly significant with high effect sizes ($r_{\text{SFB+nG \& I+G}} = 0.533$, $r_{\text{SFB+nG \& SFB+G}} = 0.525$).

In terms of the question regarding participants' satisfaction with the dialogue at this point in the interaction (Q3), the same groups as in Q2 showed significant differences. A

notable difference is observed between SFB+nG and I+G, where the effect size is once again high ($r = 0.509$). Similarly, the differences between SFB+nG and SFB+G, with an almost high effect size of $r = 0.495$, are also significant. When comparing SFB+G and I+nG, the effect size is slightly lower but still medium ($r = 0.378$), likewise to I+nG and I+G ($r = 0.426$).

6.3. Self-Assessment Questionnaires

In the following, we examine the distinctions among the four dialogue strategies concerning self-assessment questionnaires related to the user's perception of an argumentative dialogue system. Specifically, we explore their impact on overall impression, user engagement, and user trust.

The mean (M) and standard deviation (SD) were determined for each individual item within each group. For all items, the assumption of a normal distribution, based on the Shapiro–Wilk test, had to be rejected ($W = 0.574$–0.934, $p < 0.05$). Therefore, to determine the significance of the difference between the means of the four groups, we employed the non-parametric Kruskal–Wallis test [80] for k-independent samples with no specific distribution. If the Kruskal–Wallis test indicated a significant difference among the various groups, we utilized Dunn–Bonferroni as a post-hoc test to analyze which groups (pairwise) differ significantly while accounting for multiple test corrections.

6.3.1. General User Perception

The following section presents the results of a questionnaire that aligns with the ITU-T Recommendation P.851 [77]. Such questionnaires can be used to evaluate the quality of speech-based services. It consists of 32 individual items, which can be grouped into the following aspects: "information provided by the system" (IPS), "communication with the system" (COM), "user's impression of the system" (UIS), "acceptability" (ACC). Furthermore, we added seven self-formulated items addressing the aspect "argumentation" (ARG).

After the conversation with the ADS, participants were required to rate each item on a 5-point Likert scale (1 = totally disagree, 5 = totally agree). The respective questionnaire items and resulting p-value of the Kruskal–Wallis test are shown in Table 4. Moreover, for the items identified as significant, Table 5 presents the mean values M and standard deviations SD, along with the Dunn–Bonferroni corrected p-values of the pairwise comparisons. For completeness, the Appendix C includes Table A3, which presents the means and standard deviations for all items.

Table 4. Single questionnaire items concerning the user's perception of the system P.851 [77] are grouped based on the following aspects: IPS and COM. Significant differences ($\alpha = 0.05$) between the groups are indicated by a bold p-value.

Asp.	Question	p Value
IPS	1. The system has provided you the desired information.	**0.026**
	2. The system's answers and proposed solutions were clear.	**0.010**
	3. You would rate the provided information as true.	0.400
	4. The information provided by the system was incomplete *.	0.671
	$\sum$	**0.020**
COM	1. The system always understood you well.	0.680
	2. You had to concentrate in order to understand what the system expected from you. *	0.482
	3. The system's responses were well understandable.	0.255
	4. You were able to interact efficiently with the system.	0.234
	$\sum$	0.172

Table 4. *Cont.*

Asp.	Question	*p* Value
SB	1. You knew at each point of the interaction what the system expected from you.	0.056
	2. In your opinion, the system processed your responses (specifications) correctly.	0.397
	3. The system's behavior was always as expected.	0.183
	4. The system often makes mistakes in understanding you. *	0.833
	5. The system reacted appropriately.	0.195
	6. The system reacted flexibly.	0.122
	7. You were able to control the interaction in the desired way.	0.232
	8. The system reacted too slowly. *	0.166
	9. The system reacted politely.	0.165
	10. The system's responses were too long. *	0.904
	∑	0.104
DI	1. You perceived the dialogue as unnatural *.	**<0.001**
	2. It was easy to follow the flow of the dialogue.	0.138
	3. The dialogue was too long. *	0.095
	4. The course of the dialogue was smooth.	0.546
	5. You and the system could clear misunderstandings easily.	0.133
	6. You would have expected more help from the system. *	0.492
	∑	**0.005**
UIS	1. Overall, you were satisfied with the dialogue.	**0.002**
	2. The dialogue with the system was useful.	**0.005**
	3. It was easy for you to obtain the information you wanted.	0.539
	4. You have perceived the dialogue as unpleasant *.	**0.015**
	5. You felt relaxed during the dialogue.	0.653
	6. Using the system was fun.	0.117
	∑	**<0.001**
ACC	1. In the future, you would use the system again.	0.214
	2. You would would recommend the system to a friend.	0.343
	∑	0.058
ARG	1. I felt motivated by the system to discuss the topic.	**<0.001**
	2. I would rather use this system than read the arguments in an article.	**0.002**
	3. The possible options to respond to the system were sufficient.	0.536
	4. The arguments the system presented are conclusive.	0.419
	5. I felt engaged in the conversation with the system.	**0.001**
	6. The interaction with the system was confusing. *	0.320
	7. I do not like that the arguments are provided incrementally. *	0.512
	∑	**<0.001**

Items with * have to be inverted.

The category "Overall Quality" ("What is your overall impression of the system?") is not included in Table 4 as it was rated on a different 5-point Likert scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad). Our analysis shows a statistically significant ($p = 0.002$) difference between the groups.

The pairwise comparisons of the groups reveal that there is a significant difference between SFB+nG ($M = 2.36, SD = 0.93$) and SFB+G ($M = 3.54, SD = 0.78$) or I+G ($M = 3.93, SD = 1.00$) with $p = 0.033, r = 0.321$ or $(p = 0.001, r = 0.401)$, respectively. All other pairwise comparisons are not significant (p = 0.083–1.000).

Table 5. Means and standard deviations of the questionnaire items [77] in Table 4 which are significant according to the Kruskal–Wallis-test for all four groups (I+nG, SFB+nG, I+G, SFB+G) defined in Section 5. Significant differences ($\alpha = 0.05$, Dunn–Bonferroni corrected) in pairwise group comparisons are indicated by a bold p-value.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G		I+nG & SFB+nG	I+nG & I+G	I+nG & SFB+G	SFB+nG & I+G	SFB+nG & SFB+G	I+G & SFB+G
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*						
IPS	1.	3.00	1.00	2.57	1.17	**3.69**	0.86	3.57	0.76	1.00	0.402	0.706	**0.019**	**0.027**	1.00
	2.	3.00	0.93	2.43	1.16	**3.77**	0.73	3.50	1.09	1.00	0.213	1.00	**0.011**	**0.020**	1.00
	∑	3.00	0.58	2.86	0.68	**3.52**	0.60	3.41	0.50	1.00	0.264	0.503	**0.012**	**0.027**	1.00
DI	1.	3.67	1.11	3.43	1.02	**1.92**	0.95	2.00	1.18	1.00	**0.003**	**0.004**	**0.016**	**0.023**	1.00
	∑	2.82	0.49	2.59	0.53	**3.54**	0.44	3.03	0.55	1.00	**0.003**	**0.019**	**0.045**	0.054	1.00
UIS	1.	2.67	0.82	2.57	0.85	**3.93**	0.90	3.85	0.92	1.00	0.392	**0.012**	0.228	**0.006**	1.00
	2.	2.93	0.86	2.80	0.92	**3.85**	0.83	3.64	0.63	1.00	**0.018**	0.087	**0.048**	0.289	1.00
	4.	2.27	0.80	3.29	1.43	**1.77**	0.83	2.07	0.92	0.394	1.00	1.00	**0.012**	**0.017**	1.00
	∑	3.16	0.36	2.81	0.44	**3.66**	0.27	3.57	0.43	0.587	0.279	**0.030**	**0.002**	**<0.001**	1.00
ARG	1.	2.73	0.88	2.64	0.63	**4.32**	0.60	3.79	1.12	1.00	**0.001**	0.066	**<0.001**	**0.026**	1.00
	2.	2.27	1.16	2.21	1.05	3.31	1.25	**3.64**	1.01	1.00	**0.032**	**0.020**	0.124	**0.012**	1.00
	5.	2.87	1.19	2.50	1.23	**4.08**	0.76	3.93	1.14	1.00	0.061	0.102	**0.010**	**0.017**	1.00
	∑	2.67	0.59	2.68	0.48	**3.47**	0.39	3.19	0.49	1.00	**0.001**	0.184	**0.001**	**0.021**	0.750

In Table 4 the single-item analysis between the four groups does not show any significant differences regarding the aspects of communication with the system (COM), system behavior (SB) and acceptability (ACC). With regard to the information provided by the system (IPS) it can be perceived that two single items, addressing whether the provided information matched the user's request (IPS 1) and clarity of information information (IPS 2), have been rated significantly different between the groups. Table 5 provides more detailed information on which groups differ significantly after Dunn–Bonferroni correction. Both items, IPS 1 and IPS 2 were rated significantly lower for the SFB+nG compared to the gesture groups (SFB+G and I+G) with moderate effect sizes ($r_{\text{IPS}} = 0.312$–0.391). This is also perceivable in the aggregated aspect IPS, although the groups do not differ significantly regarding the truthfulness (IP 3) and completeness (IP 4) of the provided information.

Regarding the aspect of dialogue (DI), only one specific item concerning the naturalness of the dialogue (DI 1) exhibited a significant difference with moderate effect ($r_{\text{DI 1}} = 0.348$–0.357) among the four groups. In particular, Table 5 highlights that the differences between gesture and no-gesture (I+nG) are statistically very significant. This also has an impact on the aggregated aspect of DI, which shows a significant difference between SFB+nG and I+G, as well as between I+nG and the two gesture groups with moderate effect sizes ($r_{\text{DI}} = 0.362$–0.457).

Regarding the user's impression of the system (UIS), the items addressing satisfaction (UIS 1), the usefulness of the dialogue (UIS 2) and pleasantness (UIS 4, inverted) were rated with a significant difference. The pairwise comparison reveals significant differences with small to moderate effect sizes ($r = 0.216$–0.436) for UIS 1 between SFB+nG and SFB+G, for UIS 2 between the two no-gesture groups and I+G, and for UIS 4 between I+nG and SFB+G, as well as SFB+nG and SFB+G. Consequently, there is a (highly) significant difference between I+nG and SFB+G, SFB+nG and I+G, and SFB+nG and SFB+G for the aggregated aspect user's impression of the system (UIS) ($r = 0.347$–0.479).

Regarding our self-added aspect argumentation (ARG), we observed a significant difference in the single items "motivation to discuss the topic" $r_{\text{ARG 1}} = 0.364$–0.557 between I+nG and I+G, SFB+nG and SFB+G and SFB+nG and SFB+G, "preference towards reading the arguments in an article" ($r_{\text{ARG 2}} = 0.276$–0.319) and "engagement induced by the system" ($r_{\text{ARG 5}} = 0.275$–0.376). This is also reflected in the significant differences regarding the

aggregated aspect ARG with moderate effect sizes ($r = -0.480$) in the pairwise comparisons of I+nG and I+G, SFB+nG and SFB+G and SFB+nG and I+G.

6.3.2. Provided Content

Similar to the self-assessment questionnaire in the previous section, the participants were required to rate each of the six items on the conveyed content (arguments) on a 5-point Likert scale (1 = totally disagree, 5 = totally agree).

Table 6 shows the respective questionnaire items on the presented content (arguments) and the resulting p-value of the Kruskal–Wallis test. Except for one item (C6) addressing whether the user did not like any of the arguments provided, all items are statistically significant to highly significant.

Table 6. Questionnaire items regarding argument content. Significant differences are indicated by a bold p-value.

Asp.	Question	p Value
C	1. I liked the arguments suggested by the system.	**0.003**
	2. The suggested arguments fit my preference.	**<0.001**
	3. The suggested arguments were well-chosen.	**0.004**
	4. The suggested arguments were relevant.	**0.002**
	5. The system suggested too many bad arguments.*	**0.018**
	6. I did not like any of the recommended arguments. *	0.069
	$\sum$	**<0.001**

Items with * have to be inverted.

The results of the Dunn–Bonferroni corrected pairwise comparison are presented in Table 7. The means and standard deviations of all content-related questionnaire items in Table 6 are shown in Table A4 in Appendix C. Concerning the likability of the provided content (C1), a significant difference is observed between SFB+nG and I+G ($r = 0.472$) and SFB+G ($r = 0.327$). Regarding the fitness of the provided content (C2), a significant difference is noticed between SFB+nG vs I+nG ($r = 0.388$) and SFB+nG and I+G ($r = 0.496$). For the items suitable choice (C3), relevance (C4), and good argument (C5), we observed a significant difference between SFB+nG and I+G $r_{C3} = 0.407$, $r_{C4} = 0.420$, $r_{C5} = 0.406$), and SFB+nG and SFB+G ($r_{C3} = 0.415$, $r_{C4} = 0.388$, $r_{C5} = 0.320$). Regarding the aggregated aspect of the provided content (C$\sum$), the differences between I+nG and I+G ($r = 0.499$), SFB+nG and I+G ($r = 0.586$), and SFB+nG and SFB+G ($r = 0.557$) are significant.

Table 7. Means and standard deviations of the questionnaire items regarding the provided content in Table 6 which are significant according to the Kruskal–Wallis-test for all groups. Significant differences ($\alpha = 0.05$, Dunn–Bonferroni corrected) in pairwise group comparisons are indicated by a bold p-value.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G		I+nG & SFB+nG	I+nG & I+G	I+nG & SFB+G	SFB+nG & I+G	SFB+nG & SFB+G	I+G & SFB+G
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*						
C	1.	2.93	1.28	1.79	0.67	**3.69**	1.25	3.21	1.58	1.00	0.986	1.00	**0.002**	**0.046**	1.00
	2.	3.53	1.26	2.00	0.96	**3.92**	0.95	3.07	1.21	**0.011**	1.00	1.00	**0.001**	0.215	1.00
	3.	2.67	1.40	2.29	0.91	**3.69**	1.11	3.50	0.86	1.00	0.128	0.345	**0.011**	**0.039**	1.00
	4.	3.27	0.96	2.64	1.15	**4.00**	0.577	3.93	1.13	0.971	0.313	0.415	**0.007**	**0.009**	1.00
	5.	3.00	1.69	3.43	1.28	**1.85**	0.80	2.21	1.12	1.00	0.284	1.00	**0.023**	**0.040**	1.00
	$\sum$	3.12	0.47	2.42	0.39	**3.96**	0.47	3.60	0.74	0.095	**0.008**	0.424	**<0.001**	**<0.001**	0.951

6.3.3. User Trust

Table 8 illustrates the user ratings of the 11 individual items taken from the questionnaire [78]. This questionnaire was developed to measure trust in automation which was

examined to assess user trust during the interaction with the ADS. Each item was rated on a 5-point Likert scale (1 = totally disagree, 5 = totally agree). The respective questionnaire items and the resulting *p*-value of the Kruskal–Wallis test are shown in Table 8.

Table 8. Single questionnaire items of the short user engagement questionnaire Körber [78] grouped by the following aspects: understanding/predictability (UP), familiarity (F), propensity to trust (PT) and trust in automation (TA). Significant differences ($\alpha = 0.05$) between the groups are denoted by a bold *p*-value.

Asp.	Question	I+nG		SFB+nG		I+G		SFB+G		*p* Value
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
UP	1. The system state was always clear to me.	3.27	0.961	3.00	1.11	**3.38**	0.961	3.29	1.14	0.804
	2. The system reacts unpredictably.*	3.13	0.990	3.43	1.02	**2.69**	1.32	2.71	1.14	0.305
	3. I was able to understand why things happened.	3.33	1.23	3.21	1.12	**3.73**	0.832	3.64	1.28	0.078
	4. It is difficult to identify what the system will do next. *	2.93	1.10	3.86	1.03	**2.77**	1.36	2.93	0.100	0.064
	∑	3.13	0.81	2.73	0.70	**3.54**	0.72	2.68	0.93	0.051
F	1. I already know similar systems.	2.80	1.08	2.64	1.22	**2.85**	1.14	2.71	1.14	0.939
	2. I have already used similar systems.	2.60	1.08	**2.79**	1.31	2.62	1.39	2.64	1.28	0.978
	∑	2.69	1.01	2.71	1.03	**2.73**	1.17	2.68	0.93	0.997
PT	1. One should be careful with unfamiliar automated systems. *	3.67	0.82	3.98	0.54	**3.45**	0.86	3.71	0.99	0.274
	2. I rather trust a system than mistrust it.	2.73	0.88	2.79	1.12	**3.13**	0.86	2.86	0.86	0.680
	3. Automated systems generally work well.	**3.27**	0.80	2.79	0.80	3.15	0.99	2.50	0.86	0.096
	∑	2.78	0.65	2.48	0.60	**2.73**	1.17	2.55	0.61	0.099
TA	1. I trust the system.	2.87	0.92	2.64	1.00	**3.24**	0.78	3.07	1.14	0.494
	2. I can rely on the system	2.73	0.88	2.79	0.58	**3.23**	0.60	3.14	1.10	0.319
	∑	2.80	0.84	2.71	0.70	**3.24**	0.63	3.11	0.98	0.253

Items with * have to be inverted.

Neither any of the single items nor the aggregated aspects in Table 8 reached statistical significance in the differences between the four groups. Still, the mean ratings of the gesture groups were higher, especially with respect to I+G, compared to the no-gesture groups (in particular, with respect to SFB+nG). This finding suggests that users tend to trust an agent using co-speech gestures more than a static one.

6.4. User Opinion and User Interest

Furthermore, we asked the participants about their opinion on and interest in the discussion topic before ("pre-") and after ("post-") the interaction.

As shown in Table 9, the difference between the four groups regarding their "pre-opinion" on the topic of the discussion (measured on a 5-point Likert scale, where 1 represented "Totally disagree" and 5 represented "Totally agree") is insignificant. Similarly, the differences for the topic-related "pre-interest" of the participants (measured on a 5-point Likert scale before the interaction, where 1 represented "Not at all interested" and 5 represented "Very much interested") and the "post-opinion" between the groups are insignificant.

Table 9. Means and standard deviations of the user interest and opinion before (pre) and after (post) the interaction. Significant differences are indicated by a bold p-value.

Group	I+nG		SFB+nG		I+G		SFB+G		p Value
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	
pre-opinion	3.07	0.96	2.64	0.75	2.92	0.64	2.86	0.95	0.535
post-opinion	3.33	1.11	2.79	0.89	3.31	0.86	3.00	0.96	0.377
pre-interest	3.07	0.96	2.86	0.95	2.92	0.95	3.14	0.66	0.855
post-interest	2.87	0.83	2.64	1.00	3.92	0.64	4.00	1.11	**<0.001**

To determine whether there are differences between the pre- and post-ratings among the four groups, the Kruskal–Wallis test was applied. A highly significant difference can be observed for the "post-interest" with $p < 0.001$. Dunn–Bonferroni corrected pairwise comparison of the groups indicates that only the differences between the no-gesture (SFB+nG and I+nG) and gesture (SFB+G and I+G) groups are not significant (SFB+nG vs. SFB+G/I+G: $p = 0.008/0.015$; I+nG vs. SFB+G/I+G: $p = 0.023/0.043$).

Furthermore, to determine the significance of the difference between pre- and post-measurements, we utilized the non-parametric Wilcoxon signed rank test [81] for paired samples and Bonferroni corrected p-values based on a set of four comparisons. For the gesture groups, a significant difference in the user interest can be perceived before and after the interaction can be perceived (SFB+G: $p = 0.018, r = 0.315$; I+G: $p = 0.019, r = 0.311$). In the no-gesture groups, the difference between pre- and post-interest is insignificant (SFB+nG: $p = 0.546$; I+nG: $p = 0.567$). The difference between the pre- and post-opinion is insignificant for all groups (SFB+G: $p = 0.317$, I+G: $p = 0.129$, SFB+nG: $p = 0.317$, I+nG: $p = 0.157$).

6.5. User Engagement

In the following, we present the results of the 12 items introduced by the questionnaire of [79] to investigate the user's conversational engagement. The items are grouped by the following aspects: Focused attention (FA), perceived usability (PU), aesthetic appeal (AE) and Reward (RW). Again each item was rated on a 5-point Likert scale (1 = totally disagree, 5 = totally agree). The respective questionnaire items and resulting p-value of the Kruskal–Wallis test are shown in Table 10.

As evident in Table 10, except for the aspect of aesthetic appeal (AE), significant differences can be observed between the groups in all other aspects. In the Appendix C the means and standard deviations of all items in Table 10 are shown in Table A5.

In particular, all single items of the focused attention (FA) exhibit a statistically significant difference between the four groups. As shown in Table 11, for FA 1 and FA 3, there is a notable difference, particularly between SFB+nG and I+G, and SFB+nG and SGB+G, with a moderate effect size of r = 0.386–0.433. For FA 2, a significant difference after Dunn–Bonferroni correction is only observed for I+nG and SFB+G ($r = 0.386$), but a clear difference between SFB+nG and SGB+G is still notable. This is also reflected in the aggregated consideration of the aspect, which shows a highly significant difference ($p < 0.001$) between SFB+nG and the two gesture groups with strong effect sizes ($r_{SFB+nG\ \&\ SFB+G} = 0.516$; $r_{SFB+nG\ \&\ I+G} = 0.544$).

Table 10. Single questionnaire items of the short user engagement questionnaire O'Brien et al. [79]. Significant differences ($\alpha = 0.05$) between the groups are denoted by a bold p-value.

Asp.	Question	p Value
FA	1. I lost myself in this experience.	**0.002**
	2. The time I spent using the application just slipped away.	**0.009**
	3. I was absorbed in this experience.	**0.002**
	$\sum$	**<0.001**
PU	1. I felt frustrated while using the application.*	0.083
	2. I found this application confusing to use.*	**<0.001**
	3. Using this application was taxing.*	0.142
	$\sum$	**0.001**
AE	1. The application was attractive.	0.492
	2. The application was aesthetically appealing.	0.123
	3. This application appealed to my senses.	0.082
	$\sum$	0.332
RW	1. Using the application was worthwhile.	0.075
	2. My experience was rewarding.	**0.006**
	3. I felt interested in this experience.	**0.039**
	$\sum$	**0.001**

Items with * have to be inverted.

Table 11. Means and standard deviations of the questionnaire items [79] in Table 10 which are significant according to the Kruskal–Wallis test for all groups. Significant differences ($\alpha = 0.05$, Dunn–Bonferroni corrected) in pairwise group comparisons are indicated by a bold p-value.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G		I+nG &	I+nG &	I+nG &	SFB+nG &	SFB+nG &	I+G &
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	SFB+nG	I+G	SFB+G	I+G	SFB+G	SFB+G
FA	1.	2.71	1.16	1.93	0.73	3.31	1.03	**3.43**	1.16	0.432	0.943	0.511	**0.010**	**0.003**	1.00
	2.	3.07	0.97	2.50	1.35	**4.00**	0.71	3.71	0.91	1.00	0.210	1.00	**0.011**	0.094	1.00
	3.	2.87	0.92	2.43	0.85	3.62	0.65	**3.64**	0.929	1.00	0.280	0.218	**0.009**	**0.006**	1.00
	$\sum$	2.89	0.76	2.29	0.60	**3.64**	0.41	3.60	0.62	0.289	0.063	0.121	**<0.001**	**<0.001**	1.00
PU	2.	3.60	0.97	3.43	1.10	**2.23**	1.09	2.36	0.84	1.00	**0.007**	**0.021**	**0.035**	0.095	1.00
	$\sum$	2.83	0.75	2.71	0.71	**3.64**	0.55	3.56	0.62	1.00	**0.021**	0.068	**0.011**	**0.038**	1.00
RW	2.	2.87	0.99	2.43	1.16	3.46	0.66	**3.71**	0.83	1.00	0.530	0.105	0.095	**0.012**	1.00
	3.	3.75	0.66	3.36	1.51	4.31	0.48	**3.86**	0.66	1.00	0.098	1.00	**0.049**	1.00	0.546
	$\sum$	3.09	0.77	2.93	0.66	**3.82**	0.42	3.69	0.61	1.00	**0.029**	0.249	**0.004**	**0.048**	1.00

Regarding the aspect of perceived usability (PU), only one specific item concerning the naturalness of the dialogue (PU 2) exhibited a significant difference with moderate effect ($r_{\mathrm{PU}\,2}$ = 0.356–0.409) among the four groups. In particular, the differences between I+nG compared to I+G and SFB+G and are statistically significant. This also has an impact on the aggregated aspect of PU, which shows a significant difference between SFB+nG and SFB+G/I+G, as well as between I+nG and I+G with moderate effect sizes (r_{PU} = 0.340–0.361).

With regard to the aspect of reward (RW), two items, rewardingness of the experience (RW 2) and interest in the experience (RW 3), differ significantly with a moderate effect size ($r_{RW2} = 0.290, r_{RW3} = 0.302$). Furthermore, we perceive a significant difference between I+nG and I+G, SFB+nG and I+G, and SFB+nG and SFB+G concerning the aggregated aspect of reward (RW) with effect sizes r = 0.353–0.453.

7. Discussion

This section discusses the results presented in Section 6 in relation to the hypotheses introduced in Section 5.

7.1. Validation of H1

Regarding the first hypothesis that co-speech gestures positively influence the perception of the provided content (particularly for the SFB+G group) (H1), we examine the turnwise ratings of Q1 in Section 6.2, the self-assessment questionnaire in Section 6.3.2, and the general impressions in Section 6.3.1.

We anticipated that the interest groups would outperform the SFB groups, as their arguments were tailored to the user's interests and requests. This expectation is reflected in the turnwise ratings of Q1, where SFB+nG performed the worst with a slightly below-neutral average. The difference between I+nG and SFB+nG, although not significant, is noticeable. Surprisingly, I+nG performs worse than SFB+G, even though SFB+G's argument selection does not match user preferences. The comparison between SFB+nG and SFB+G, which share the same argument selection strategy but differ in presentation, shows a strong impact on content perception. Although I+G performs better than SFB+G, there is no significant difference, indicating that arguments selected by the SFB-breaking policy and presented with co-speech gestures are perceived significantly better.

This observation is confirmed by the results in Section 6.3.2, which show a strong statistically significant difference in self-assessment ratings of the provided content post-interaction. Arguments selected based on the interest strategy tend to perform better within their respective no-gesture or gesture groups (I+nG vs. SFB+nG, I+G vs. SFB+G). Notably, for item C2 (the suggested arguments fitted my preference), only the differences between I+nG and SFB+nG, and I+G and SFB+nG were significant, but not between SFB+G and the interest groups. Thus, both in turnwise and post-evaluation, arguments were perceived more positively when presented with co-speech gestures, even with the same selection strategy. This is particularly evident in the comparison between SFB+G and SFB+nG, where arguments presented with co-speech gestures were rated as more likable (C1), well-chosen (C3), relevant (C4), and less bad (C5). Aggregated data show SFB+nG rating I+G and SFB+G significantly better, and I+G rated significantly better than I+nG, despite the same argument selection strategy.

This aligns with the results in the aspect "information provided by the system" (IPS) in Section 6.3.1. It shows that the subjective impressions of the provided information, such as desire (IPS1) and clarity (IPS2), differ significantly between SFB+nG and I+G/SFB+G. However, co-speech gestures do not appear to affect the objective impressions of truthfulness (IPS3) and completeness (IPS4).

Consequently, we conclude that although arguments based on the interest selection strategy are preferred, there is a significant difference when this content is presented with co-speech gestures. Co-speech gestures enhance the perception of the provided content, especially for arguments selected by the SFB-breaking strategy, even if they contradict the user's opinion. This confirms our hypothesis H1.

7.2. Validation of H2

To confirm our hypothesis H2 that participants in the two SFB-breaking (interest) groups do not show a significant difference in the efficiency in breaking the user's SFB, we discuss the results of Section 6.1.

These results address the overall SFB dimensions and can be similarly observed for individual clusters. As shown in the box plots in Figure 7, SFB+G and SFB+nG, as well as I+G and I+nG, show almost no difference in all overall SFB dimensions. In contrast, the differences between I+nG vs. SFB+nG, I+nG vs. SFB+G, I+G vs. SFB+nG, and I+G vs. SFB+G are highly significant for all dimensions. The high effect size in Reflective User Engagement (RUE) and the medium to high effect size in True Knowledge (TK) support our assumption that co-speech gestures do not influence the user's SFB vector or

its change during the interaction. Thus, they do not affect the user's reflective engagement, aligning with the optimization of the SFB-breaking strategy for maximizing RUE and TK. Additionally, the medium to high effect in Personal Relevance and the predominantly medium effect in False Knowledge indicate significant differences across all dimensions, demonstrating the strategy's overall success, as shown by Aicher et al. [16].

Moreover, this finding aligns with Murali et al. [82], who studied tailoring virtual exercise coaches for different cultures by manipulating appearance and argumentation. They found that adapting argumentation, rather than appearance and non-verbal behavior alone, was essential for effectiveness. Similarly, in our study, co-speech gestures alone do not influence the user's SFB.

Thus, as expected, there are no statistically significant differences between groups with the same argument selection strategy, confirming our hypothesis H2. The SFB-breaking strategy is effective in breaking the user's SFB, in contrast to the interest selection strategy.

7.3. Validation of H3

Concerning the third hypothesis (H3), which posits that participants across all four groups exhibit no significant difference in their opinion formation, we discuss the findings outlined in Section 6.4.

The initial opinions on the discussion topic were found to be comparable across all four groups, as indicated by the non-significant differences in pre-opinion ratings. This similarity ensures a fair comparison across all experimental conditions. This observation holds true for the post-opinion ratings as well. The results reveal that there is no significant difference in post-opinion ratings among the four groups. No discernible difference is apparent between the gesture and no-gesture groups. However, a marginal trend in the ratings for the interest groups suggests a slight inclination to support the major claim, while the ratings for the SFB-breaking groups show a trend toward neutrality.

This observation remains consistent when comparing pre- and post-opinion across all groups. In conclusion, the consistent non-significant differences in pre- and post-opinion ratings across all groups suggest that co-speech gestures did not have a discernible impact on participants' substantive opinions. From this, we deduce that although co-speech gestures have a positive effect on the perception of the conveyed content (see Section 7.1), they do not manipulate users' opinion forming during the interaction and thus, confirm our hypothesis H3. Therefore, we conclude that user opinion is not significantly influenced by not personalized, pre-defined co-speech gestures, which suggests that they appear to be an effective means of making the interaction conversationally more engaging. This, of course, needs to be further investigated in each subsequent step of individualization and personalization of the virtual agent to ensure that the user emerges from their self-imposed filter bubble through reflective engagement with the content and not the influence of personalized non-verbal behavior.

7.4. Validation of H4

The findings in Section 6.3.3 are discussed regarding the fourth hypothesis (H4), which states that participants across all four groups show no significant difference in their self-assessed trust.

No statistically significant differences were found between the four groups for both individual items and their aggregation. However, Table 8 reveals certain trends: mean ratings for the gesture groups, especially I+G, are slightly higher than for the no-gesture groups. SFB+nG performs the least favorably, particularly when compared to SFB+G, suggesting the argument selection policy might influence trust, especially without co-speech gestures. Notably, generic co-speech gestures alone do not significantly impact user trust. Even when SFB-breaking groups received arguments not matching their expectations, their trust in the agent remained unaffected, indicating users accept autonomous agent decisions as they would from a human partner.

These results highlight the importance of transparent and unbiased opinion formation, showing that the co-speech gestures used do not inherently lead to user manipulation. Therefore, the lack of significant differences in self-assessed user trust confirms hypothesis H4.

7.5. Validation of H5

The findings regarding the fifth hypothesis (H5)—stating that differences in engagement and general perception are less within gesture and no-gesture groups (SFB+G vs. I+G, SFB+nG vs. I+nG) compared to those with similar argument choice strategies (SFB+G vs. SFB+nG, I+G vs. I+nG)—are discussed based on Sections 6.3.1 and 6.5 and the turnwise ratings of Q2, Q3 in Section 6.2.

Users in the gesture groups showed higher engagement levels, as detailed in Table 10 and the pairwise comparisons in Table 11. The co-speech gesture system significantly impacted user engagement [79], with medium effect sizes in "focused attention" (FA), "perceived usability" (PU), and "reward" (RW), but no significant difference in "aesthetic appeal" (AE) due to consistent avatar and GUI aesthetics across groups. No significant difference in "aesthetic appeal" (AE) is observed, which is expected since the attractiveness/appeal of the application, especially the avatar and GUI design, did not vary between the groups.

The positive impact of co-speech gestures is further supported by significant differences in single items and aggregated aspect ARG in Table 4. Notable differences are observed in ARG 1 ("I felt motivated by the system to discuss the topic"), ARG 2 ("I would rather use this system than read the arguments in an article."), and ARG 5 ("I felt engaged in the conversation with the system").

Regarding turnwise ratings of presentation (Q2), gesture groups clearly outperformed no-gesture groups, with medium to high effect sizes indicating that co-speech gestures significantly enhance the users' motivation and engagement.

Section 6.3 shows a significant difference in "Overall Quality" (OQ) between gesture groups (SFB+G, I+G) and SFB+nG. In terms of dialogue satisfaction (Q3), gesture groups significantly outperformed no-gesture groups, with SFB+nG performing worst. Interestingly, the difference between SFB+nG and I+nG is small despite the user interest-tailored argument selection for I+G. However, differences between no-gesture and gesture groups are much greater, with SFB+G and I+G rated significantly better. The most substantial effects are observed between SFB+nG and I+G, followed by SFB+nG and SFB+G, and I+nG and I+G, indicating that co-speech gestures significantly influence overall user impression.

This is further supported by the user's impression of the system (UIS) in Table 4, where gesture groups rated much better in satisfaction (UIS 1), dialogue usefulness (UIS 2), and pleasantness (UIS 4). This is particularly evident between SFB+nG and SFB+G, also reflected in the aggregated UIS aspect. The naturalness of dialogue (DI 1) was rated significantly better for the gesture groups. Due to the study setting, with intermediate ratings after each interaction turn, the dialogue cannot be compared to a fluent conversation. This likely affects the impression of naturalness, even though participants were asked to exclude it from their ratings. However, since this condition applies to all groups, comparisons and conclusions remain valid.

In conclusion, our results show that the user engagement and the user's general perception of the agent are significantly higher for gesture groups compared to no-gesture groups. The differences between groups with similar argument choice strategies (SFB+nG vs. SFB+G, I+nG vs. I+G) are insignificant, confirming hypothesis H5.

These findings suggest that co-speech gestures enhance the perception of the agent and positively influence user engagement. This effect is particularly notable with the SFB-breaking policy, as co-speech gestures seem to mitigate negative perceptions of content not aligning with user requests.

8. Limitations

This study has several limitations that future research could address. Firstly, with only 56 participants discussing a single topic ("Marriage is an outdated institution") sourced from one dataset, the scalability of our findings to other topics and larger user groups remains uncertain. We selected this topic because its dataset met our criteria: it was sufficiently extensive, provided a balanced range of pro and con arguments, maintained high argument quality, and offered depth in reasoning. Nevertheless, future studies should investigate a broader scope of topics and involve a larger number of participants.

Secondly, our study compared a static virtual agent to one with pre-defined, motion-captured gestures from `Vuppetmaster` (https://www.charamel.com/products/vuppetmaster, accessed on 22 July 2024). These gestures were not customized for specific arguments or individual user responses but were instead consistent across users to ensure comparability. Future studies should investigate the effects of personalized and content-specific gestures.

Additionally, we focused solely on co-speech gestures and did not incorporate the agent's listening behavior during user responses. Future research should model responsive listening to create more natural interactions.

Our study also examined only one aspect of non-verbal communication. Future work should explore the full range of non-verbal cues, including posture, gaze, facial expressions, and emotions, in both speech and listening behaviors of virtual agents. Additionally, the cultural backgrounds of users should be considered, encompassing both the content and the non-verbal behavior of the agent.

Furthermore, even though it would be intriguing to address users directly about their current estimated SFB, this might be perceived as criticism, leading to reluctance, especially if users disagree they are caught in an SFB. However, it is possible to gauge users' openness to new arguments indirectly through the SFB-dimension Reflective User Engagement (RUE) and their behavior in accepting suggested arguments. The SFB provides a continuous description of the user's SFB level via the SFB-vector, which is constantly updated and represents the probability of the user being stuck in an SFB. This vector serves as the metric against which the user's SFB is compared.

Likewise, the user-agent interaction might seem limited because users cannot introduce counterarguments, but this choice was intentional. The objective is to present users with pro and con arguments neutrally, allowing exploration without direct persuasion. According to [34], users tend to defend their views and confrontation with opposing arguments leads to cognitive dissonance [83] and defensive attitudes [84], making rejection likely. However, we aim to facilitate a more natural exchange in future iterations. We plan to explore how users can introduce their arguments without reinforcing their SFBs. To achieve this, we propose dynamically searching for relevant arguments via state-of-the-art argument search engines in real-time, providing arguments for both viewpoints to enrich the discourse.

Finally, improvements are needed in perceived usability (PU), particularly regarding Automatic Speech Recognition errors and system responses when users are not understood correctly (COM 1, COM 2, COM 4).

As our results clearly indicate, co-speech gesture groups received significantly higher ratings in overall quality compared to the static agent. To further enhance system performance, it is crucial to address these limitations and tailor the agent's behavior to better align with user expectations and non-verbal cues.

9. Conclusions and Outlook

Previous research suggests that the non-verbal behavior of virtual and embodied agents significantly influences user motivation and actions [66]. With the growing importance of the social web, understanding the impact of multimodal agent behavior on interpersonal communication, especially in argumentation, is crucial [19].

Therefore, in this work, we aim to develop a system that helps users break their self-imposed filter bubbles (SFBs) by investigating the role of co-speech gestures and examining

their significant contributions to this objective. In particular, we explored how a virtual agent's co-speech gestures affect user perception, interest, trust, opinion formation, and engagement in argumentative dialogues under two different argument selection strategies. After presenting the models for these strategies, we outlined the agent's static and co-speech gesture behaviors. A laboratory experiment with 56 participants was then conducted, and the results were analyzed.

Our results confirmed several hypotheses. As we expected, we found no significant difference in the effectiveness of breaking the user's SFB between the two SFB-breaking (Interest) groups. However, there is a significant difference between the SFB-breaking and Interest groups, highlighting the influence of the respective argument selection strategies. Specifically, the SFB-breaking selection effectively breaks the user's SFB and promotes critical scrutiny of information on the given topic. Additionally, participants across all four groups showed no significant difference in opinion-building and self-assessed trust. Although co-speech gestures positively influenced content perception, particularly in the SFB+G group, they did not affect users' opinion formation or trust. As co-speech gestures do not manipulate users' opinion formation or user trust during the interaction, they are suitable for maintaining user motivation and conversational engagement, supporting strategies that foster reflective user engagement.

Beyond the significant positive influence of co-speech gestures on content perception, participants in the gesture and no-gesture groups exhibited less disparity within their respective groups (SFB+G vs. I+G, SFB+nG vs. I+nG) regarding engagement and general perception of the agent, compared to groups with similar argument choices (SFB+G vs. SFB+nG, I+G vs. I+nG). The argument selection strategy has a slight but not significant influence on the overall perception of the agent and user engagement. In contrast, the gesture groups showed a significant enhancement in conversational engagement, interest, and overall perception of the virtual agent, even when gestures were not tailored to argument content or user response. Since co-speech gestures do not influence users' opinion formation or trust, they are an effective approach to enhancing user engagement in argumentative dialogues.

In future research, we aim to address the mentioned limitations and further explore the potential of adapting non-verbal agent behavior and individually generated gestures to enhance interactions within argumentative dialogue systems. Our next step involves analyzing the influence of more complex gesture selection, tailored to argument content. Specifically, we will investigate how individualized "expressivity dimensions" [62] of the agent impact user perception, interest, opinion, and trust with respect to the agent's personality. Additionally, we plan to model gestures to better fit the provided content, including integrating multiple gestures in a single turn and modeling the agent's listening behavior. In particular, we aim to explore the possibility of generating gestures in real time during interactions.

To ensure scalability, we will test our results and approaches on datasets covering diverse controversial topics and with larger participant groups. We also plan to integrate new arguments using advanced argument mining techniques and search engines to better tailor the content. This should enhance the system's flexibility and naturalness, improving reflective and conversational engagement. Additionally, we will explore various real-time gesture generation methods for argumentative dialogues [67,85,86].

In conclusion, our findings enhance the understanding of multimodal user-agent interactions and offer valuable insights for developing systems that encourage thoughtful reflection and engage users without manipulation. Thus, co-speech gestures help to mitigate the negative perception if the provided content does not align with the user's request, and therefore, support fostering a well-founded opinion-building and sustaining an interaction to disrupt/break the user's SFB. This work represents a step forward in improving multimodal interactions with argumentative virtual agents by enhancing both reflective engagement through tailored SFB-breaking arguments and conversational engagement through integrated co-speech gestures used by an embodied conversational agent.

Author Contributions: Conceptualization, A.A., Y.M., K.Y. and S.U.; methodology, A.A. and S.U.; software, A.A.; validation, A.A.; formal analysis, A.A.; investigation, A.A; resources, W.M., K.Y. and E.A.; data curation, A.A. and Y.M.; writing—original draft preparation, A.A.; writing—review and editing, A.A., S.U. and E.A.; visualization, A.A.; supervision, W.M., K.Y. and S.U.; project administration, W.M. and K.Y.; funding acquisition, A.A., W.M., Y.M., K.Y. and E.A. All authors have read and agreed to the published version of the manuscript.

Funding: 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) and JST PRESTO, Grant no. JPMJPR2039.

Institutional Review Board Statement: This study was conducted in accordance with and approved by the Ethical Review Committee for Research Involving Human Subjects at Nara Institute of Science and Technology (Approval No.: 2020-I-1-1, approved on 15 March 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Acceptability (questionnaire aspect)
ADS	Argumentative dialogue system
AE	Aesthetic appeal (questionnaire aspect)
ARG	Argumentation (questionnaire aspect)
C	Content (questionnaire aspect)
DI	Dialogue (questionnaire aspect)
COM	Communication with the system
ECA	Embodied conversational agent
ELM	Elaboration Likelihood Model
F	Familiarity (questionnaire aspect)
FA	Focused attention (questionnaire aspect)
FK	False Knowledge
G	Gesture
I	Interest
IPS	Information provided by the system (questionnaire aspect)
nG	No-gesture
PR	Personal Relevance
PT	propensity to trust (questionnaire aspect)
PU	Perceived usability (questionnaire aspect)
RUE	Reflective User Engagement
RW	Reward (questionnaire aspect)
SB	System behavior (questionnaire aspect)
SFB	Self-imposed Filter Bubble
TA	Trust in automation (questionnaire aspect)
TK	True Knowledge
UIS	User's impression of the system (questionnaire aspect)
UP	Understanding/predictability (questionnaire aspect)
VA	Virtual agent

Appendix A. Materials and Methods

Appendix A.1. Subtreewise User Interest and Weights

Let B_{i_k} be the subtree with root node i_k (where i denotes a unique identifier) belonging to cluster c; $|B_{i_k}|$ the total number of descendants of i. If the user has visited $|B_{i_k,v}|$ of these descendants, the user interest with regard to subtree B_{i_k} is defined as:

$$I(B_{i_k}) = \frac{|B_{i_k,v}|}{|B_{i_k}|} \tag{A1}$$

where $I(B_{i_k}) \in [0,1]$. $I(B_{i_k}) = 0$ means that the user has not requested any further information on argument node i and consequently, is not interested in cluster k. $I(B_{i_k}) = 1$ means that all available information has been requested and thus the user has the largest possible interest in k with respect to the presented subtree.

Intuitively, the overall cluster interest could be determined from the individual interest values $I(B_{i_k})$ (see Equation (A1)) by taking their average [71]. However, averaging would treat all interest values equally, without considering differences in subtree sizes. Consequently, small subtrees would be over-represented, while larger ones would be under-represented. To address this issue, we introduce a weight that accounts for different subtree sizes. To determine the overall cluster interest of k, a subtree B_{i_k} is weighted with:

$$\omega_{n,B_{i_k}} = \frac{|B_{i_k}|}{\sum_{r_k} |B_{r_k}|}, \tag{A2}$$

where $\omega_{n,B_{i_k}} \in [0,1]$. $\omega_{n,B_{i_k}}$ represents the relative size of a subtree B_{i_k} by considering the number of arguments contained in all subtrees B_{r_k} with root nodes k belonging to cluster k.

Considering hierarchical argumentation structures, arguments on different levels differ in information detail. Specifically, an argument component located at the beginning of a branch (lower level) is more general than one deeper down (higher level). The latter provides in-depth information and contains many more details. Therefore, hierarchical weighting is used to incorporate the varying levels of argument depth into the Interest Model. It is assumed that a user who is highly interested will ask for more detailed information. Thus, subtrees starting at lower levels will be assigned larger weight values than subtrees with root nodes closer to the Major Claim. As the weights must have ascending values, the argument level is divided by the sum of the levels of the current branch.

Let node i be a descendant of the branch root node j. The maximum depth of the corresponding argument branch B_j is denoted by d_{max,B_j}, and the level of i is denoted by d_{i,B_j}. We define the weight $\omega_{d,B_{i_k}}$ for subtree B_{i_k} with root node i belonging to cluster k as follows:

$$\omega_{d,B_{i_k}} = \frac{d_{i,B_j}}{\sum_{l=1}^{d_{max,B_j}-1} l}, \tag{A3}$$

with ω_d in $[0,1]$. Note, that for leaf nodes, no succeeding arguments exist and thus no further information is available. Consequently, the upper limit of the sum in the denominator of Equation (A3) ends one level above the leaf node.

Appendix A.2. Examplary Calculation of Overall User Interest

To determine the user interest in the cluster "Children (Ch)," three visited subtrees with root nodes i_{Ch} in Claim 1, Premise 1, and Premise 3 have to be considered. As only children of Claim 1 were visited, it follows that $I(B_{i_{Ch}}) = \frac{4}{6} = \frac{2}{3}$, whereas it is 0 for the other two subtrees. Since there are three subtrees that contain 6+1+1 = 8 nodes, $\omega_{n,B_{i_{Ch}}} = \frac{6}{8} = \frac{3}{4}$ for Claim 1 and $\frac{1}{8}$ for the other two. By dividing the respective level (1 for Claim 1 and

2 for the rest) by $d_{max,B_j} = 3$, we obtain $\omega_{d,B_{i_c}} = \frac{1}{3}$ and $\frac{2}{3}$. Substituting these values into Equation (1), we find that the user interest in the cluster *Ch* to be:

$$I_{Ch} = \frac{\frac{2}{3}\frac{3}{4}\frac{1}{3} + 0 + 0}{\frac{3}{4}\frac{1}{3} + \frac{1}{8}\frac{2}{3} + \frac{1}{8}\frac{2}{3}} = 0.4, \tag{A4}$$

indicating a moderate user interest in this cluster.

Appendix A.3. SFB Model Dimensions

Appendix A.3.1. Reflective User Engagement (RUE)

The elaboration likelihood model (ELM) [72], a well-established framework in persuasion research, proposes that attitude change occurs through two information processing modes—central vs. peripheral. Westerwick et al. [87] assert that central route processing involves careful, thorough engagement with information, wherein users reflect upon it, integrate it with preexisting cognitions, and assimilate it into their cognitive network. Conversely, lacking motivation and ability for such cognitive effort may lead to peripheral processing, and diminishing scrutiny of information content. Consequently, peripheral processing heightens the probability of users being ensnared in their SFB.

Reflective User Engagement (RUE) denotes the manifestation of critical thinking and open-mindedness exhibited by users when exploring controversial topics [15]. Critical thinking and open-mindedness are frequently advocated strategies to counteract various biases [33,88,89]. Given the definition of confirmation bias and SFBs, which inherently oppose reflective engagement, it follows that *RUE* significantly influences the user's SFB vector. Hence, a higher *RUE* concerning a specific cluster diminishes the likelihood of users being entrenched in a cluster SFB.

Building upon our previous work Aicher et al. [14,15], we propose a RUE calculation incorporating argument polarity and quantity (pertaining to a cluster) heard by a user. Since RUE is defined as a user's interest in scrutinizing arguments and exploring divergent views, we quantify this through user actions, such as seeking more information on either side (pro/con) of the discussion topic. Our model requires users' awareness of the number of available arguments, depicted through corresponding visualization in the graphical user interface, and thereby taking into account a deliberate omission rather than oversight of unheard arguments.

Consequently, a higher RUE is associated with a higher balanced number of heard pro and con arguments. We compute the RUE dependent on respective clusters k. The RUE increases if the number of heard pro and con arguments is balanced or/and the more unknown arguments are heard. To mitigate potential data-related biases in cluster k, we introduce the characteristic function (A5).

$$\mathbb{1}_{p_{k,v}} = \begin{cases} 1, & \text{if } \exists \text{ visited pro/con pairs } \wedge\, s_{k,v,\rho} \leq s_{k,v,\bar{\rho}} + \min\{2, (s_{k,a} - s_{k,v})\} \\ 1, & \text{if } \nexists \text{ visited pro/con pairs } \wedge\, s_{k,v,\rho} < s_{k,v,\bar{\rho}} \\ 1, & \text{if no pro/con pairs exist } \wedge\, s_{k,v,\rho} \leq s_{k,v,\bar{\rho}} + \min\{2, (s_{k,a} - s_{k,v})\} \\ 0, & \text{if } s_{k,v,\rho} > s_{k,v,\bar{\rho}} + \min\{2, (s_{k,a} - s_{k,v})\} \\ 0, & \text{if } \nexists \text{ visited pro/con pairs } \wedge\, s_{k,v,\rho} \geq s_{k,v,\bar{\rho}} \end{cases} . \tag{A5}$$

Throughout Section 3.4, s_k denotes the number of single arguments belonging to cluster k and p_k denote pro/con pairs. No direct connection between the arguments is required here; the arguments just need to belong to the same cluster k and have different polarities. ($p_k = s_{k,\rho} \wedge s_{k,\bar{\rho}}$) of the respective cluster k. The index a denotes all elements in this cluster, v denotes visited and thus heard arguments, ρ denotes an argument's polarity corresponding to the user's point of view, and $\bar{\rho}$ denotes an argument's polarity contradicting the user's point of view. If users claim indifference, $s_{k,v} = s_{k,v,\rho}$.

Equation (A5) considers whether at least one pro/con pair has been heard, allowing additional heard single arguments to be considered, with limitations imposed on cases

where users hear additional arguments contradicting their own viewpoint or only a few arguments supporting it. Consequently, a balanced exploration or exploration of opposing views is rewarded more, as it demands greater user effort.

If additional single arguments exist ($s_k \geqq 0$), we define these singles with $s_{k,w}$ in Equation (A6), distinguishing between three cases. The first describes when the number of visited single arguments $s_{k,v}$ is smaller than the total number of single arguments $s_{k,a}$ in cluster k. The second considers when the user explores more singles (here, only arguments that are still unknown to the user are considered. If the user states to already know the argument $\gamma_k := 0$.) than pairs, in line with their point of view, weighed with a downsizing factor $\gamma_k \in (0,1)$. The third considers when the user explores more opposing singles than pairs, indicating higher RUE, with an additional factor $\beta_{1,k}, \beta_{2,k} \in (0,1]$, which can be adapted to the reward exploration of opposing arguments more.

$$s_{k,w} = \begin{cases} \beta_{1,k}\frac{s_{k,v}}{s_{k,a}}, & \text{if } s_{k,v} \leqq s_{k,a} \\ \gamma_k, & \text{if } s_{k,v} > s_{k,a} \ \wedge s_{k,v,\rho} \geqq s_{k,v,\bar{\rho}} \\ \beta_{2,k}\frac{s_{k,v,\bar{p}} - p_{k,v}}{s_{k,a,\bar{p}} - p_{k,v}} & \text{if } s_{k,v} > s_{k,a} \ \wedge s_{k,v,\rho} < s_{k,v,\bar{\rho}} \end{cases}. \tag{A6}$$

In Equation (A6), $s_{k,v}$ denotes visited single arguments either in line with or opposing the user's viewpoint. $s_{k,a}$ represents all existing singles in a cluster, and the numerator of the term $s_{k,v,\bar{p}}$ indicates the number of visited arguments opposing the user's viewpoint in cluster k, subtracting the already heard pairs $p_{k,v}$ from it. The denominator consists of all arguments opposing the user's viewpoint $s_{k,a,\bar{p}}$ subtracted by $p_{k,v}$. Thus, Equation (A6) acknowledges that exploration of opposing views counteracts the SFB, exerting a higher impact than exploration of arguments reinforcing the user's viewpoint.

The resulting RUE component r_k of cluster k is determined by:

$$rue_k = \alpha_k \frac{|p_{k,v}|}{|p_{k,a}|} + \mathbb{1}_{p_{k,v}}(1-\alpha_k)s_{k,w}, \tag{A7}$$

where $rue_k \in [0,1]$. In Equation (A7), visited pairs $p_{k,v}$ are weighted with a factor α_k, and all single arguments with $(1-\alpha_k)$, allowing for a balanced exploration reward. If no pro/con pairs exist in cluster k ($|p_{k,a}| = 0$), we define $\alpha_k := 0$; $\frac{|p_{k,v}|}{|p_{k,a}|} := 0$.

Appendix A.3.2. Personal Relevance (PR)

According to Westerwick et al. [87], the choice between the central and peripheral information processing modes furthermore depends on the individual user motivation, i.e., personal relevance. Thus, the Personal Relevance (PR) displays another dimension in our SFB Model. It refers to the user's individual assessment of how relevant a cluster is with regard to the topic of the discussion. The greater the relevance a cluster holds for a user, the stronger their inclination to delve into the corresponding arguments associated with it. As this is impossible to ascertain through implicit methods, the Personal Relevance is explicitly queried within the dialogue when transitioning to a new cluster, with respect to the previous cluster (see Figure 3). Thus, each cluster is assigned a certain rating, here on a 5-point Likert-scale. For instance, the user could rate the statement "This aspect is personally relevant to me in the discussion of {topic}" for each cluster: 5 = Strongly agree, 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly disagree.

By normalizing the obtained rating, we obtain the personal relevance pr of the respective cluster k:

$$pr_k = \frac{\text{user rating}}{5}, \tag{A8}$$

with $pr_k \in [0,1]$. Thus, we conclude that the bigger the Personal Relevance regarding a certain cluster k, the higher the user's motivation to explore arguments belonging to k.

Appendix A.3.3. True and False Knowledge

Besides the previously mentioned dimensions, the ability, i.e., preexisting knowledge [87], is crucial for a user to process information via the central route and thus, thoroughly scrutinize it according to the ELM model. Building upon this argumentation, we consider (preexisting) knowledge by distinguishing two correlated dimensions: True Knowledge (TK) and False Knowledge (FK).

True Knowledge serves as a measure of information gain and is defined as the new information the user is provided with by talking to the ADS. It can be determined by comparing the total information provided by the system and the information, which is already known to the user. For its determination, the user is required to provide feedback on each known argument. As we want the user to explore as much information as possible, a high *TK* increases the chance to explore other aspects and viewpoints. Thus, the bigger the *TK* of the users, the more unlikely they find themselves in an SFB.

Consequently, the True Knowledge tk_k concerning a cluster k is defined as the relation of the number of arguments belonging to k that the user listens to during the interaction and the total number of arguments belonging to k (n_k). As we want to distinguish between the preexisting knowledge of the user and the newly gained one through the interaction, we furthermore define the initial True Knowledge $tk_{k,i}$ as the relation of the number of arguments the user states to already know ($n_{k,v,known}$) and n_k. It follows:

$$tk_k = \frac{n_{k,v}}{n_k}, \tag{A9}$$

$$tk_{k,i} = \frac{n_{k,v,known}}{n_k}, \tag{A10}$$

with $tk_k, tk_{k,i} \in [0,1]$. $n_{k,v}$ denotes the number of all visited arguments ($n_{k,v,known} \leq n_{k,v}$).

The dimension False Knowledge. Regarding the terminology, please note that the term False Knowledge was chosen to facilitate a simplified three-dimensional representation, wherein the dimensions of True Knowledge and False Knowledge are merged into a single dimension of Knowledge. This choice is intended solely for the purpose of simplified illustration as the actual calculation occurs within a four-dimensional space. Without loss of generality, the information stored in the system's database is defined as factually accurate, thereby classifying information contradicting it as wrong. This pertains to inaccurate information held by a user regarding a specific topic. When a user possesses false beliefs about specific clusters, it increases the probability of being caught in an SFB and fosters reluctance toward conflicting information and viewpoints. Likewise, to True Knowledge, False Knowledge is determined by an explicit user statement. Here, the users indicate that they consider an argument to be factually incorrect.

To display False Knowledge fk_k regarding arguments belonging to cluster k in a similar range $(0,1]$, the following inverse relation is defined:

$$fk_k = \frac{1}{1+\theta_k n} \quad \forall n \in \mathbb{N}_0, \tag{A11}$$

where n denotes the number of instances where the user is stated to have contradicting information and $\theta \in (0,\infty)$ displays a weighting factor that can be chosen accordingly. Without loss of generality, we choose $\theta_k = 0.5 \; \forall \; k$ as a starting point. As it might be useful to adjust θ_k according to the cluster sizes, we chose a ten-dimensional $\vec{\theta}$ instead of θ, as the herein discussed dataset consists of 10 clusters. In the case of $n = 0$, we define $fk_k := 1$, and thus, no False Knowledge regarding cluster k is present. Therefore, the probability for the user to be open-minded towards presented arguments is higher.

Appendix B. ADS Architecture

Example Interaction

In Tables A1 and A2, we present segments of exemplary dialogues with the ADS. Both illustrate an artificial interaction between the system and the user regarding the topic *Marriage is an outdated institution*, sourced from the *Debatabase* (https://idebate.org/debatabase, accessed on 23 July 2021).

Table A1. Artificial dialogue between the ADS and the user on the topic "*Marriage is an outdated institution*". The arguments are selected according to the Interest selection strategy. The actor, utterance, respective moves, and updated Interest components of cluster k are shown.

Actor	Utterance	Move	Interest Comp.
System	...		
User	Could you please provide an attacking argument?	why_{con}	I_{loss}, $\lvert B_{i_k,v}\rvert$
System	One could argue that marriage is an important institution to religious people.		I_{loss}, $\lvert B_{i_k,v}\rvert$
User	I understand, but could you elaborate this further?	why_{pro}	I_{loss}, $\lvert B_{i_k,v}\rvert$
System	This is supported by the fact, that marriage is an integral part of most major religions, particularly Christianity, which encompasses over 40% of the population of the UK according to the British Social Attitudes Survey of 2007.		
User	Yes, I totally agree.	$prefer$	I_{loss}
System	I see. Well, as you prefer the presented argument, I would like to draw your attention to another aspect that might be of interest to you...		

Table A2. Artificial dialogue between the ADS and the user on the topic "*Marriage is an outdated institution*". The arguments are selected according to the SFB-breaking policy. The actor, utterance, respective moves, and updated SFB dimensions of cluster k are shown.

Actor	Utterance	Move	SFB Dim.
System	...		
User	Could you please provide an attacking argument?	why_{con}	rue_k
System	Rather than focusing solely on an attacking argument, I propose we also consider the opposing view. One could argue that religion as a whole is becoming less significant, leading to a reduced importance of marriage.	why_{pro_sfb}	tk_k
User	I understand, but could you provide additional evidence?	why_{pro}	rue_k
System	On the pro side, it is worth mentioning that the number of people in the UK identifying as religious has declined by 20% in the last 20 years, as indicated by the British Social Attitudes Survey of 2007.		tk_k
User	Impressive, but isn't this statistic outdated by now? I don't believe it holds true anymore.	$false$	fk_k
System	I understand your skepticism, but if you allow me, I'd like to provide a more detailed explanation...		

Material reproduced from www.iedebate.org with the permission of the International Debating Education Association (Copyright© 2005 International Debate Education Association. All Rights Reserved.)

While in Table A1 the arguments are selected according to the Interest selection strategy, the selection in Table A2 follows the SFB-breaking policy.

Appendix C. Results

For completeness, the following tables extend the results presented in Section 6. Table A3 presents the means and standard deviations for all items of Table 4.

Table A3. Means and standard deviations of the questionnaire items regarding the provided content in Table 4 for all groups.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
IPS	1.	3.00	1.00	2.57	1.17	**3.69**	0.86	3.57	0.76
	2.	3.00	0.93	2.43	1.16	**3.77**	0.73	3.50	1.09
	3.	3.33	0.62	3.50	0.65	**3.77**	0.60	3.50	0.76
	4.	3.33	0.90	3.07	1.07	2.99	1.07	**2.93**	0.92
	∑	3.00	0.58	2.86	0.68	**3.52**	0.60	3.41	0.50
COM	1.	2.13	1.06	**2.36**	0.93	2.31	1.03	2.00	0.96
	2.	3.77	0.59	3.93	1.00	**3.62**	0.65	3.89	0.68
	3.	3.20	0.78	3.21	1.12	**3.77**	0.60	3.14	1.10
	4.	2.20	0.86	2.43	0.94	**2.92**	0.86	2.36	1.08
	∑	2.40	0.51	2.52	0.69	**2.85**	0.55	2.38	0.80
SB	1.	2.87	0.99	3.14	1.10	**3.69**	0.86	2.64	1.15
	2.	2.53	0.99	2.93	1.10	**3.08**	0.83	2.71	0.91
	3.	2.47	0.83	2.50	0.86	**3.15**	1.07	2.43	1.02
	4.	3.87	0.92	3.64	0.93	**3.54**	1.05	3.79	0.89
	5.	3.27	1.13	3.50	0.65	**3.31**	1.18	**3.64**	0.94
	6.	2.60	1.30	2.71	1.20	3.31	1.18	**3.43**	0.94
	7.	2.53	1.25	2.29	1.07	**3.08**	0.95	2.57	1.02
	8.	4.00	0.85	4.07	1.07	**3.23**	1.17	3.17	0.91
	9.	3.00	0.84	3.29	1.20	3.45	1.47	**3.86**	0.95
	10.	3.00	0.78	3.14	1.23	**2.77**	1.09	3.00	0.96
	∑	2.60	0.53	2.77	0.40	**3.06**	0.43	2.88	0.51
DI	1.	3.67	1.11	3.43	1.02	**1.92**	0.95	2.00	1.18
	2.	3.40	0.83	3.14	1.03	**3.46**	0.97	2.64	1.15
	3.	3.13	0.99	2.57	1.16	**2.15**	0.80	2.71	1.20
	4.	2.87	1.19	2.86	1.03	**3.38**	1.04	2.93	1.14
	5.	2.80	1.01	3.00	0.78	**3.38**	0.77	2.64	1.01
	6.	3.33	0.82	3.29	0.83	**2.92**	0.86	3.36	0.75
	∑	2.82	0.49	2.59	0.53	**3.54**	0.44	3.03	0.55
UIS	1.	2.67	0.82	2.57	0.85	**3.93**	0.90	3.85	0.92
	2.	2.93	0.86	2.80	0.92	**3.85**	0.83	3.64	0.63
	3.	2.73	0.80	2.64	0.84	**3.15**	0.99	2.93	1.07
	4.	2.27	0.80	3.29	1.43	**1.77**	0.83	2.07	0.92
	5.	3.13	0.74	2.64	1.01	3.45	0.69	**3.50**	0.76
	6.	3.87	0.92	3.36	0.84	3.86	1.12	**4.00**	0.88
	∑	3.16	0.36	2.81	0.44	**3.66**	0.27	3.57	0.43
ARG	1.	2.73	0.88	2.64	0.63	**4.32**	0.60	3.79	1.12
	2.	2.27	1.16	2.21	1.05	3.31	1.25	**3.64**	1.01
	3.	2.80	0.94	3.14	1.03	**3.31**	0.75	3.00	0.88
	4.	2.93	0.70	2.79	0.80	**3.23**	0.83	2.79	0.80
	5.	2.87	1.19	2.50	1.23	**4.08**	0.76	3.93	1.14
	6.	3.47	1.19	3.43	0.76	**2.77**	1.01	3.43	1.16
	7.	3.46	0.83	3.09	0.92	**3.08**	0.76	3.36	0.75
	∑	2.67	0.59	2.68	0.48	**3.47**	0.39	3.19	0.49

Furthermore, Table A4 presents the means and standard deviations for all items of Table 6 in Section 6.3.2.

Table A4. Means and standard deviations of the questionnaire items regarding the provided content in Table 6 for all groups.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
C	1.	2.93	1.28	1.79	0.67	**3.69**	1.25	3.21	1.58
	2.	3.53	1.26	2.00	0.96	**3.92**	0.95	3.07	1.21
	3.	2.67	1.40	2.29	0.91	**3.69**	1.11	3.50	0.86
	4.	3.27	0.96	2.64	1.15	**4.00**	0.577	3.93	1.13
	5.	3.00	1.69	3.43	1.28	**1.85**	0.80	2.21	1.12
	6.	2.67	1.29	2.71	1.33	**1.69**	0.63	1.93	0.83
	∑	3.12	0.47	2.42	0.39	**3.96**	0.47	3.60	0.74

Additionally, Table A5 provides the means and standard deviations for all items listed in Table 10 from Section 6.5.

Table A5. Means and standard deviations of the questionnaire items [79] in Table 10 for all groups.

Asp.	No.	I+nG		SFB+nG		I+G		SFB+G	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
FA	1.	2.71	1.16	1.93	0.73	3.31	1.03	**3.43**	1.16
	2.	3.07	0.97	2.50	1.35	**4.00**	0.71	3.71	0.91
	3.	2.87	0.92	2.43	0.85	3.62	0.65	**3.64**	0.929
	∑	2.89	0.76	2.29	0.60	**3.64**	0.41	3.60	0.62
PU	1.	3.00	1.13	3.07	1.07	2.54	0.87	**2.21**	0.80
	2.	3.60	0.97	3.43	1.10	2.23	1.09	2.36	0.84
	3.	3.27	1.03	3.00	0.88	2.54	0.88	**2.50**	1.02
	∑	2.83	0.75	2.71	0.71	**3.64**	0.55	3.56	0.62
AE	1.	**3.33**	1.06	2.86	1.17	3.23	1.30	3.29	1.07
	2.	3.07	0.70	3.14	0.86	**3.69**	0.49	3.43	0.85
	3.	3.00	0.76	3.02	0.88	**3.46**	0.65	3.21	0.70
	∑	3.20	0.65	3.00	0.82	**3.51**	0.59	3.31	0.53
RW	1.	2.87	1.06	3.00	0.88	**3.69**	0.75	3.50	0.76
	2.	2.87	0.99	2.43	1.16	3.46	0.66	3.71	0.83
	3.	3.75	0.66	3.36	1.51	4.31	0.48	3.86	0.663
	∑	3.09	0.77	2.93	0.66	3.82	0.42	3.69	0.61

References

1. Liu, Y.; Mohammadi, G.; Song, Y.; Johal, W. Speech-Based Gesture Generation for Robots and Embodied Agents: A Scoping Review. In Proceedings of the 9th International Conference on human–agent Interaction (HAI '21), Virtual Event, 9–11 November 2021; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 31–38. [CrossRef]
2. Wang, I.; Ruiz, J. Examining the Use of Nonverbal Communication in Virtual Agents. *Int. J. Hum.-Comput. Interact.* **2021**, *37*, 1–26. [CrossRef]
3. Nickerson, R.S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **1998**, *2*, 175–220.
4. Ekström, A.G.; Niehorster, D.C.; Olsson, E.J. Self-imposed filter bubbles: Selective attention and exposure in online search. *Comput. Hum. Behav. Rep.* **2022**, *7*, 100226. [CrossRef]
5. Aicher, A.; Minker, W.; Ultes, S. Towards Modelling Self-imposed Filter Bubbles in Argumentative Dialogue Systems. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022.
6. Quattrociocchi, W.; Scala, A.; Sunstein, C.R. Echo Chambers on Facebook. SSRN 2795110; 2016; pp. 1–15. Available online: https://www.researchgate.net/publication/323980520_Echo_Chambers_on_Facebook (accessed on 29 March 2024).
7. Anand, B.N. The US media's problems are much bigger than fake news and filter bubbles. In *Domestic Extremism*; Greenhaven Publishing: New York, NY, USA, 2021; p. 138.

8. Donkers, T.; Ziegler, J. The Dual Echo Chamber: Modeling Social Media Polarization for Interventional Recommending. In Proceedings of the 15th ACM Conference on Recommender Systems (RecSys '21), Amsterdam, The Netherlands, 27 September–1 October 2021; Association for Computing Machinery (ACM): New York, NY, USA, 2021; pp. 12–22. [CrossRef]
9. Pariser, E. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*; Penguin: New York, NY, USA, 2011.
10. Aicher, A.; Gerstenlauer, N.; Feustel, I.; Minker, W.; Ultes, S. Towards Building a Spoken Dialogue System for Argument Exploration. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 1234–1241.
11. Barnidge, M.; Peacock, C.; Kim, B.; Kim, Y.; Xenos, M.A. Networks and selective avoidance: How social media networks influence unfriending and other avoidance behaviors. *Soc. Sci. Comput. Rev.* **2023**, *41*, 1017–1038.
12. Terren, L.T.L.; Borge-Bravo, R.B.B.R. Echo chambers on social media: A systematic review of the literature. *Rev. Commun. Res.* **2021**, *9*, 99–118.
13. Ross Arguedas, A.; Robertson, C.; Fletcher, R.; Nielsen, R. Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review. 2022. Available online: https://reutersinstitute.politics.ox.ac.uk/echo-chambers-filter-bubbles-and-polarisation-literature-review (accessed on 19 January 2022).
14. 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 '23), Eindhoven, The Netherlands, 19–21 July 2023; Association for Computing Machinery (ACM): New York, NY, USA, 2023. [CrossRef]
15. Aicher, A.; Minker, W.; Ultes, S. Determination of Reflective User Engagement in Argumentative Dialogue Systems. In Proceedings of the Workshop on Computational Models of Natural Argument (CMNA), Online, 2–3 September 2021.
16. Aicher, A.; Kornmueller, D.; Matsuda, Y.; Ultes, S.; Minker, W.; Yasumoto, K. Towards Breaking the Self-imposed Filter Bubble in Argumentative Dialogues. In Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue, Prague, Czechia, 11–15 September 2023; Stoyanchev, S., Joty, S., Schlangen, D., Dusek, O., Kennington, C., Alikhani, M., Eds.; pp. 593–604. [CrossRef]
17. 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 '23), New York, NY, USA, 19–22 September 2023; Association for Computing Machinery (ACM): New York, NY, USA, 2023. [CrossRef]
18. Miao, F.; Kozlenkova, I.V.; Wang, H.; Xie, T.; Palmatier, R.W. An Emerging Theory of Avatar Marketing. *J. Mark.* **2022**, *86*, 67–90, [CrossRef]
19. Blount, T.; Millard, D.E.; Weal, M.J. On the Role of Avatars in Argumentation. In Proceedings of the 2015 Workshop on Narrative & Hypertext (NHT '15), Guzelyurt, Northern Cyprus, 1 September 2015; Association for Computing Machinery (ACM): New York, NY, USA, 2015; pp. 17–19. [CrossRef]
20. Aicher, A.; Gerstenlauer, N.; Minker, W.; Ultes, S. User Interest Modelling in Argumentative Dialogue Systems. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 127–136.
21. Aicher, A.; Matsuda, Y.; Yasumoto, K.; Minker, W.; André, E.; Ultes, S. Exploring the Impact of Non-Verbal Virtual Agent Behavior on User Engagement in Argumentative Dialogues. In Proceedings of the HAI 2024, Swansea, UK, 24–27 November 2024; *Unpublished work, accepted for publication*.
22. Dwyer, C.P.; Hogan, M.J.; Stewart, I. *The Promotion of Critical Thinking Skills through Argument Mapping*; Nova Science Publishers, Inc.: Hauppauge, NY, USA, 2011; Journal of Information Technology Education: Research, Volume 22
23. Dwyer, C.P. An Evaluative Review of Barriers to Critical Thinking in Educational and Real-World Settings. *J. Intell.* **2023**, *11*, 105. [CrossRef]
24. Lucas, P. Critical Reflection. What Do We Really Mean; In Proceedings of the Australian Collaborative Education Network (ACEN) National Conference Deakin University, Geelong, Australia, 29 October–2 November 2012.
25. Rüütmann, T. Development of Critical Thinking and Reflection. In The Challenges of the Digital Transformation in Education: Proceedings of the 21st International Conference on Interactive Collaborative Learning (ICL2018), Budapest, Hungary, 27–29 September 2019; Auer, M.E., Tsiatsos, T., Eds.; Springer International Publishing: Cham, Switzerland; pp. 895–906.
26. Whitaker, L.; Reimer, E. Students' conceptualisations of critical reflection. *Soc. Work Educ.* **2017**, *36*, 946–958, [CrossRef]
27. Indrašienė, V.; Jegelevičienė, V.; Merfeldaitė, O.; Penkauskienė, D.; Pivorienė, J.; Railienė, A.; Sadauskas, J. Critical Reflection in Students' Critical Thinking Teaching and Learning Experiences. *Sustainability* **2023**, *15*, 13500. [CrossRef]
28. Makhene, A. Argumentation: A Methodology to Facilitate Critical Thinking. *Int. J. Nurs. Educ. Scholarsh.* **2017**, *14*, 20160030. [CrossRef]
29. Allahverdyan, A.E.; Galstyan, A. Opinion dynamics with confirmation bias. *PLoS ONE* **2014**, *9*, e99557.
30. Jones, M.; Sugden, R. Positive confirmation bias in the acquisition of information. *Theory Decis.* **2001**, *50*, 59–99.
31. Kappes, A.; Harvey, A.H.; Lohrenz, T.; Montague, P.R.; Sharot, T. Confirmation bias in the utilization of others' opinion strength. *Nat. Neurosci.* **2020**, *23*, 130–137.
32. Huang, H.H.; Hsu, J.S.C.; Ku, C.Y. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decis. Support Syst.* **2012**, *53*, 438–447.
33. Schwind, C.; Buder, J. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Comput. Hum. Behav.* **2012**, *28*, 2280–2290.

34. Paul, R.W. Critical and reflective thinking: A philosophical perspective. In *Dimensions of Thinking and Cognitive Instruction*; North Central Regional USA: Ames, IA, USA, 1990; pp. 445–494.
35. Mason, M. Critical thinking and learning. *Educ. Philos. Theory* **2007**, *39*, 339–349.
36. Gelter, H. Why is reflective thinking uncommon. *Reflective Pract.* **2003**, *4*, 337–344. [CrossRef]
37. Maloney, E.A.; Retanal, F. Higher math anxious people have a lower need for cognition and are less reflective in their thinking. *Acta Psychol.* **2020**, *202*, 102939.
38. Del Vicario, M.; Scala, A.; Caldarelli, G.; Stanley, H.E.; Quattrociocchi, W. Modeling confirmation bias and polarization. *Sci. Rep.* **2017**, *7*, 40391. [CrossRef]
39. Villarroel, C.; Felton, M.; Garcia-Mila, M. Arguing against confirmation bias: The effect of argumentative discourse goals on the use of disconfirming evidence in written argument. *Int. J. Educ. Res.* **2016**, *79*, 167–179.
40. Slonim, N.; Bilu, Y.; Alzate, C.; Bar-Haim, R.; Bogin, B.; Bonin, F.; Choshen, L.; Cohen-Karlik, E.; Dankin, L.; Edelstein, L. An autonomous debating system. *Nature* **2021**, *591*, 379–384. [CrossRef]
41. Rosenfeld, A.; Kraus, S. Strategical Argumentative Agent for Human Persuasion. In Proceedings of the Twenty-Second European Conference on Artificial Intelligence (ECAI'16), The Hague, The Netherlands, 29 August–2 September 2016; pp. 320–328. [CrossRef]
42. Rakshit, G.; Bowden, K.K.; Reed, L.; Misra, A.; Walker, M.A. Debbie, the Debate Bot of the Future. In Proceedings of the Advanced Social Interaction with Agents—8th International Workshop on Spoken Dialog Systems, Farmington, PA, USA, 6–9 June 2017; pp. 45–52.
43. 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, Brussels, Belgium, 1 November 2018; pp. 121–130. [CrossRef]
44. Hadoux, E.; Hunter, A.; Polberg, S. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *Argum. Comput.* **2022**, *14*, 1–53. [CrossRef]
45. Chalaguine, L.A.; Hunter, A. A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns. In Proceedings of the COMMA, Perugia, Italy, 4–11 September 2020. [CrossRef]
46. Chalaguine, L.; Hunter, A. Addressing Popular Concerns Regarding COVID-19 Vaccination with Natural Language Argumentation Dialogues. In Proceedings of the Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Prague, Czech Republic, 21–24 September 2021; Vejnarová, J., Wilson, N., Eds.; Springer International Publishing: Cham, Switzerland; pp. 59–73.
47. Aicher, A.; Rach, N.; Minker, W.; Ultes, S. Opinion building based on the argumentative dialogue system BEA. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 307–318.
48. Mancini, M.; Hartmann, B.; Pelachaud, C. Non-verbal behaviors expressivity and their representation. *PF-Star Rep.* **2007**, *3*, 1–9.
49. Kendon, A. *Gesture: Visible Action as Utterance*; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]
50. Mcneill, D. Hand and Mind: What Gestures Reveal About Thought. *Bibliovault OAI Repos. Univ. Chic. Press* **1994**, *37*, 203–209. [CrossRef]
51. Deichler, A.; Wang, S.; Alexanderson, S.; Beskow, J. Learning to generate pointing gestures in situated embodied conversational agents. *Front. Robot. AI* **2023**, *10*, 1110534.
52. Hasegawa, D.; Kaneko, N.; Shirakawa, S.; Sakuta, H.; Sumi, K. Evaluation of speech-to-gesture generation using bi-directional LSTM network. In Proceedings of the 18th International Conference on Intelligent Virtual Agents, Sydney, NSW, Australia, 5–8 November 2018; pp. 79–86.
53. Endrass, B.; Damian, I.; Huber, P.; Rehm, M.; André, E. Generating Culture-Specific Gestures for Virtual Agent Dialogs. In Proceedings of the 10th International Conference on Intelligent Virtual Agents, (IVA 2010), Philadelphia, PA, USA, 20–22 September 2010; Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 329–335.
54. Wolfert, P.; Robinson, N.; Belpaeme, T. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *IEEE Trans. Hum.-Mach. Syst.* **2022**, *52*, 379–389. [CrossRef]
55. Ravenet, B.; Pelachaud, C.; Clavel, C.; Marsella, S. Automating the Production of Communicative Gestures in Embodied Characters. *Front. Psychol.* **2018**, *9*, 1144. [CrossRef]
56. Watson-Smith, H.; Marcon Swadel, F.; Hutton, J.; Marcon, K.; Sagar, M.; Blackett, S.; Rebeiro, T.; Biddle, T.; Wu, T. Real Time Gesturing in Embodied Agents for Dynamic Content Creation. In Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '23), London, UK, 29 May–2 June 2023; pp. 3068–3069.
57. Yazdian, P.J.; Chen, M.; Lim, A. Gesture2Vec: Clustering Gestures using Representation Learning Methods for Co-speech Gesture Generation. In Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Kyoto, Japan, 23–27 October 2022; pp. 3100–3107. [CrossRef]
58. Binder, J.F. Establishing conversational engagement and being effective: The role of body movement in mediated communication. *Acta Psychol.* **2023**, *233*, 103840. [CrossRef]
59. Binder, J.F.; Cebula, K.; Metwally, S.; Vernon, M.; Atkin, C.; Mitra, S. Conversational engagement and mobile technology use. *Comput. Hum. Behav.* **2019**, *99*, 66–75. [CrossRef]
60. Olafsson, S.; O'Leary, T.K.; Bickmore, T.W. Motivating Health Behavior Change with Humorous Virtual Agents. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), Virtual Event, 20–22 October 2020; Association for Computing Machinery (ACM): New York, NY, USA, 2020; Volume 42, pp. 1–8. [CrossRef]

61. Neff, M.; Wang, Y.; Abbott, R.; Walker, M. Evaluating the Effect of Gesture and Language on Personality Perception in Conversational Agents. In Proceedings of the 10th International Conference on Intelligent Virtual Agents (IVA 2010), Philadelphia, PA, USA, 20–22 September 2010; Allbeck, J., Badler, N., Bickmore, T., Pelachaud, C., Safonova, A., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 222–235.
62. Pelachaud, C. Studies on gesture expressivity for a virtual agent. *Speech Commun.* **2009**, *51*, 630–639 [CrossRef]
63. Gebhard, P.; Baur, T.; Damian, I.; Mehlmann, G.; Wagner, J.; Andre, E. Exploring Interaction Strategies for Virtual Characters to Induce Stress in Simulated Job Interviews. May 2014; Volume 1. Available online: https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/deliver/index/docId/45456/file/45456.pdf (accessed on 15 March 2024)
64. Sinatra, A.M.; Pollard, K.A.; Files, B.T.; Oiknine, A.H.; Ericson, M.; Khooshabeh, P. Social fidelity in virtual agents: Impacts on presence and learning. *Comput. Hum. Behav.* **2021**, *114*, 106562. [CrossRef]
65. de Wit, J.; Brandse, A.; Krahmer, E.; Vogt, P. Varied Human-Like Gestures for Social Robots: Investigating the Effects on Children's Engagement and Language Learning. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI '20), Cambridge, UK, 23–26 March 2020; Association for Computing Machinery (ACM): New York, NY, USA, 2020; pp. 359–367. [CrossRef]
66. Gratch, J.; Wang, N.; Gerten, J.; Fast, E.; Duffy, R. Creating Rapport with Virtual Agents. In Proceedings of the 7th International Conference on Intelligent Virtual Agents (IVA 2007), Paris, France, 17–19 September 2007; Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 125–138.
67. He, Y.; Pereira, A.; Kucherenko, T. Evaluating Data-Driven Co-Speech Gestures of Embodied Conversational Agents through Real-Time Interaction. In Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents (IVA '22), Faro, Portugal, 6–9 September 2022; Association for Computing Machinery (ACM): New York, NY, USA, 2022. [CrossRef]
68. Stab, C.; Gurevych, I. Annotating Argument Components and Relations in Persuasive Essays. In Proceedings of the COLING, Dublin, Ireland, 23–29 August 2014; pp. 1501–1510.
69. Daxenberger, J.; Schiller, B.; Stahlhut, C.; Kaiser, E.; Gurevych, I. Argumentext: Argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum* **2020**, *20*, 115–121.
70. Gauch, S.; Speretta, M.; Chandramouli, A.; Micarelli, A. User profiles for personalized information access. In *The Adaptive Web. Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4321, pp. 54–89.
71. Yi, J.; Zhang, Y.; Yin, M.; Zhao, X. A novel user-interest model based on mixed measure. *J. Phys. Conf. Ser.* **2017**, *887*, 012061.
72. Petty, R.E.; Briñol, P.; Priester, J.R. Mass media attitude change: Implications of the elaboration likelihood model of persuasion. In *Media Effects*; Routledge: London, UK, 2009; pp. 141–180.
73. Bechhofer, S. OWL: Web ontology language. In *Encyclopedia of Database Systems*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 2008–2009.
74. Ekman, P. An argument for basic emotions. *Cogn. Emot.* **1992**, *6*, 169–200.
75. 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.* **2022**, *242*, 108318.
76. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. 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, Minneapolis, MN, USA, 2–7 June 2019; Volume 1: Long and Short Papers; pp. 4171–4186. [CrossRef]
77. Recommendation P.851, I.T.R. *Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems (11/2003)*; International Telecommunication Union: Geneva, Switzerland, November 2003.
78. Körber, M. 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*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 13–30.
79. 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. Hum.-Comput. Stud.* **2018**, *112*, 28–39.
80. Kruskal, W.H.; Wallis, W.A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621.
81. Wilcoxon, F. Individual Comparisons by Ranking Methods. In *Biometrics Bulletin*; International Biometric Society, Wiley: Hoboken, NJ, USA, 1945; Volume 1.
82. Murali, P.; Shamekhi, A.; Parmar, D.; Bickmore, T. Argumentation is More Important than Appearance for Designing Culturally Tailored Virtual Agents. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20), Auckland, New Zealand, 9–13 May 2020; pp. 1940–1942.
83. 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.* **2009**, *135*, 555.
84. Harmon-Jones, E. Cognitive Dissonance and Experienced Negative Affect: Evidence that Dissonance Increases Experienced Negative Affect Even in the Absence of Aversive Consequences. *Personal. Soc. Psychol. Bull.* **2000**, *26*, 1490–1501. [CrossRef]
85. Krome, N.; Kopp, S. Towards Real-time Co-speech Gesture Generation in Online Interaction in Social XR. In Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (IVA '23), Würzburg, Germany, 19–22 September 2023; Association for Computing Machinery (ACM): New York, NY, USA, 2023. [CrossRef]

86. Nyatsanga, S.; Kucherenko, T.; Ahuja, C.; Henter, G.E.; Neff, M. A Comprehensive Review of Data-Driven Co-Speech Gesture Generation. In *Computer Graphics Forum*; Wiley Online Library: Hoboken, NJ, USA, 2023; Volume 42, pp. 569–596.
87. Westerwick, A.; Johnson, B.K.; Knobloch-Westerwick, S. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Commun. Monogr.* **2017**, *84*, 343–364. [CrossRef]
88. Alsharif, H.; Symons, J. Open-mindedness as a Corrective Virtue. *Philosophy* **2021**, *96*, 73–97. [CrossRef]
89. Macpherson, R.; Stanovich, K.E. Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learn. Individ. Differ.* **2007**, *17*, 115–127. [CrossRef]