Towards Breaking the Self-imposed Filter Bubble in Argumentative Dialogues

Annalena Aicher
Ulm University, Germany
NAIST, Japan
annalena.aicher@uni-ulm.de

Daniel Kornmüller
Wolfgang Minker
Ulm University, Germany

Stefan Ultes
University of Bamberg, Germany

Yuki Matsuda
Keiichi Yasumoto
NAIST, Japan

Abstract

Human users tend to selectively ignore information that contradicts their pre-existing beliefs or opinions in their process of information seeking. These "self-imposed filter bubbles" (SFB) pose a significant challenge for cooperative argumentative dialogue systems aiming to build an unbiased opinion and a better understanding of the topic at hand.

To address this issue, we develop a strategy for overcoming users' SFB within the course of the interaction. By continuously modeling the user's position in relation to the SFB, we are able to identify the respective arguments which maximize the probability to get outside the SFB and present them to the user. We implemented this approach in an argumentative dialogue system and evaluated in a laboratory user study with 60 participants to show its validity and applicability. The findings suggest that the strategy was successful in breaking users' SFBs and promoting a more reflective and comprehensive discussion of the topic.

1 Introduction

Spoken dialogue systems are getting increasingly popular, especially as they enable easy access to requested information from online sources, such as search engines or social media platforms. Specifically, with regard to more complex interactions, two important phenomena can be observed that can result in information bias.

On the one hand, due to filter algorithms, information content is selected based on previous online behavior, which leads to cultural/ideological bubbles, the so-called "Filter Bubbles" (Pariser, 2011). On the other hand, Nickerson (1998) points out that users who are confronted with controversial topics tend to focus on a "biased subset of sources that repeat or strengthen an already established or convenient opinion." This user behavior leads to the so-called "Self-imposed Filter Bubbles" (SFB) (Ekström et al., 2022; Aicher et al., 2022b) and "echo chambers" (Quattrociocchi et al., 2016; Anand, 2021; Donkers and Ziegler, 2021). Both are manifestations of "confirmation bias", a term typically used in psychological literature. These phenomena are mutually dependent according to Lee (2019) as the SFB is reinforced and perpetuated due to algorithmic filters delivering content aligned with presumed interests based on search histories. Moreover, Bakshy et al. (2015) claim that studies have shown that individual choice has even more of an effect on exposure to differing perspectives than "algorithmic curation". In this paper, we focus on the second phenomenon, namely the user's SFB regarding a controversial topic during the interaction with an argumentative dialogue system (ADS). Building upon the work of Aicher et al. (2022b, 2023), we model the user's SFB using the following four main dimensions: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*.

The concept of *RUE* encapsulates the user's critical thinking, building upon the definition established in our prior work (Aicher et al., 2021a). On the other hand, *PR* pertains to the individual user's assessment of the significance of subtopics, further on called "clusters", in relation to the overarching topic of discussion. *True Knowledge (TK)* is characterized as the information already possessed by the user on a particular topic. Conversely, *False Knowledge (FK)* entails the user's false beliefs and misinformation on the respective topic. Based upon these dimensions we have the ability to construct a model for assessing the likelihood of a user being caught within an SFB. In order to achieve this, we ascertain the user's position along these four dimensions and consistently update it throughout the course of the dialogue. Building upon SFB-Model we 1) introduce a rule-based system policy to break the user's SFB during an ongoing interaction and 2) validate our policy in a laboratory study by comparing it to a user-interest-driven system

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 593–604
September 11–15, 2023. ©2023 Association for Computational Linguistics

policy.

The remainder of this paper is as follows: Section 2 gives an overview of related literature, followed by a description of the underlying SFB-Model and our proposed rule-based SFB-breaking policy in Section 3. Section 4 discusses an exemplary integration of our model/policy in an ADS, which is evaluated in a laboratory study described in Section 5. Section 6 covers the respective study results, followed by a discussion of the former and study limitations in Sections 6 and 8. We close with a conclusion and a brief discussion of future work in Section 9.

2 Related Work

In the following, we provide a brief overview of the existing literature on the main aspects of the work presented herein, *Confirmation Bias and Self-imposed Filter Bubbles* and *Argumentative Dialogue Systems*.

2.1 Confirmation Bias and Self-imposed Filter Bubbles

As previously pointed out, a central issue in the process of opinion building is the phenomenon known as "confirmation bias". This bias refers to the tendency of users to seek or interpret evidence in ways that align with their existing beliefs, expectations, or hypotheses (Nickerson, 1998). Given our goal of achieving a well-founded and unbiased exploration of information, we are determined to counteract the user's inclination to focus solely on information that confirms their preexisting beliefs (Allahverdyan and Galstyan, 2014).

To address this challenge, Huang et al. (2012) propose the utilization of computer-mediated counter-arguments within decision-making processes. Additionally, Schwind and Buder (2012) consider preference-inconsistent recommendations as a promising approach to stimulate critical thinking. However, given our cooperative approach and the objective of maintaining the user's motivation to explore arguments without bias, introducing an excessive number of counter-arguments could potentially lead to undesirable negative emotional consequences, such as annoyance and confusion (Huang et al., 2012).

In order to identify a means of mitigating these consequences, it is crucial to consider how a genuine and profound critical reflection can be stimulated. When users engage in critical thinking in a *weak sense*, this implies contemplating positions that differ from their own (Mason, 2007), but often involves a tendency to defend their own viewpoint without thorough introspection (Paul, 1990). Critical thinking in a *strong sense* involves reflecting on one's own opinions as well, which aligns with our objective. However, the substantial energy and effort (Gelter, 2003) required for this robust critical reflection are frequently lacking due to a deficiency in individuals' inherent *need for cognition* (Maloney and Retanal, 2020). Given users' tendency to defend their own views (Paul, 1990), a system that confronts them with opposing viewpoints might not necessarily foster critical reflection; on the contrary, it could lead to a reinforcement of their existing stance. Hence, there is a need for an intelligent system capable of adjusting the frequency, timing, and selection of counter-arguments (Huang et al., 2012). To the best of our knowledge we are the first to provide such a system, which integrates a model to determine the user's Self-imposed Filter Bubble (SFB) and adapts its strategy accordingly. This adaptation aims to identify the most suitable arguments and still maintaining the user's interest, ensuring a well-balanced exploration of viewpoints.

In contrast to Del Vicario et al. (2017), who study online social debates and try to mathematically model the related polarization dynamics, we define a model for this "seeking or interpreting of evidence in ways that are partial to existing beliefs, expectations or a hypothesis in hand" (Nickerson, 1998) consisting of four dimensions building upon our previous work (Aicher et al., 2022b, 2023). The respective dimensions are based on a well-established framework in persuasion research, the " Elaboration Likelihood Model" (ELM) (Petty et al., 2009).

2.2 Argumentative Dialogue Systems

Within this paper we define a system policy aiming to help users overcome their SFBs in a cooperative argumentative dialogue. Argumentative dialogue systems (ADS) enable users to engage in information-seeking and to explore pro and con arguments on a controversial topic by accessing large-scale argumentation structures and assist in a well-founded opinion building (Waheed et al., 2021; Aicher et al., 2021b,a, 2023). The "ability to engage in argumentation is essential for humans to understand new problems, to perform scientific

reasoning, to express, to clarify and to defend their opinions in their daily lives" (Palau and Moens, 2009) and thus, enables to reflect controversial topics critically. A consensual dialogue is much more likely to resolve diverging perspectives on evidence and repair incorrect, partial, and subjective readings of evidence than a persuasive one (Villarroel et al., 2016). Hence, it is crucial for the argumentative dialogue system, in which our SFB-Model is embedded, that it does not try to persuade or win a debate against a user.

Most approaches to human-machine argumentation utilize different models to structure the interaction and are embedded in a competitive, persuasive scenario. For instance, Slonim et al. (2021) introduced the IBM Debater, which is an autonomous debating system that can engage in a competitive debate with humans via natural language. Another speech-based approach was introduced by Rosenfeld and Kraus (2016), presenting a system based on weighted Bipolar Argumentation Frameworks (wBAG). Arguing chatbots such as Debbie (Rakshit et al., 2017) and Dave (Le et al., 2018) interact via text with the user. A menu-based framework that incorporates the beliefs and concerns of the opponent was presented by Hadoux et al. (2022). In the same line, Chalaguine and Hunter (2020) used a previously crowd-sourced argument graph and considered the concerns of the user to persuade them. Another introduced persuasive prototype chatbot is tailored to convince users to vaccinate against COVID-19 using computational models of argument (Chalaguine and Hunter, 2021). As pointed out in Subsection 2.1 in contrast to those persuasive approaches we chose collaborative exploration of arguments, enabling users to express their preferences and thus providing a more suitable basis than the previously mentioned, competitive ADS.

3 Self-imposed Filter Bubble Model

In the following section we will give a short overview on the SFB-Model we adapted to and its respective dimensions. This serves as a basis for our system's SFB-breaking policy introduced in Subsection 3.3.

3.1 SFB-Model Dimensions

We adapted the SFB-Model introduced by Aicher et al. (2022b) which is motivated by the "Elaboration Likelihood Model" (ELM) (Petty et al., 2009). As already mentioned, it incorporates of four dimensions, which span a four-dimensional space to describe the user's SFB: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*.

The *Reflective User Engagement (RUE)* describes the critical-thinking and open-mindedness demonstrated by the user. It takes into account the polarity and number of heard arguments. This can be mapped onto the request for more information, either on the pro or con side of the topic of the discussion. Thus, it measures how balanced the user is exploring a topic. The *RUE* has first been introduced by Aicher et al. (2021a), to whose work we refer to for details of its calculation.

The *Personal Relevance (PR)* 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 (PR)* is explicitly queried within the dialogue when transitioning to a new cluster, with respect to the previous cluster.

The *True Knowledge (TK)* serves as a measure for the information gain and is defined as the new information the user is provided with by talking to the system. 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. For each cluster, this number of known arguments is subtracted from the total number of arguments heard within the cluster. 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.

The concept of "False Knowledge (FK)[1]" pertains to inaccurate information held by a user regarding a specific topic. When a user possesses false beliefs about specific clusters, it increases

[1]Regarding the terminology, please note that the term "False Knowledge" was chosen to facilitate a simplified three-dimensional representation, wherein the dimensions of "True" and "False Knowledge" are merged into the 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.

the probability to be caught in an SFB and fosters reluctance toward conflicting information and viewpoints. Likewise to the "True Knowledge", the "False Knowledge" is determined by the user indicating that they consider an argument to be factually incorrect.

3.2 SFB-Model

Argumentative discussions are complex and consist of a lot of different cluster, which contain arguments referring to the same content-related aspects. For each of these clusters, a corresponding SFB vector $\overrightarrow{sfb_k} = (pr_k, r_k, tk_k, fk_k)^T,\ k \in \mathbb{N}$ is defined, contributing to the overall SFB vector $\overrightarrow{SFB_k}$ for the entire discussion topic. It is important to differentiate between the SFB and the SFB-vector of a user (refer to Figure 1). The SFB-vector is conceptualized as a vector originating from the coordinate system's origin and terminating at the user's position in the four-dimensional space. The SFB, on the other hand, constitutes the region within the four-dimensional space that signifies a specific probability of users to be caught in their SFB. Figure 1 presents an illustration[2] of two positions of this vector, and the respective SFB (dark blue geometric shape) for a single cluster. As it is very difficult to establish precise boundaries of the SFB, we establish a probability denoting a user's position within or outside the SFB. A short SFB-vector (dashed red arrow) corresponds to a high probability of the user to be caught within the SFB. Conversely, a large SFB vector (continuous green arrow) that extends further beyond the SFB diminishes the likelihood of the user to be caught in the SFB. The overall SFB vector $\overrightarrow{SFB} = (PR, RUE, TK, FK)^T$, consists of the overall cluster values for each dimension, derived from a weighted mean calculation (Aicher et al., 2023).

3.3 SFB-breaking policy

Building upon the model described in Subsection 3.2, we propose 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 5, 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; medium probability = random average).

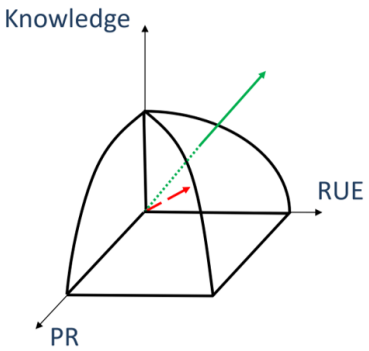

Figure 1: Schematic sketch of a clusterwise SFB-vector and SFB for a cluster k. The probability of an SFB is very high in proximity to the origin and/or when a dimension approaches a value close to zero. As a four-dimensional space is challenging to visualize, we consolidate the dimensions of *TK* and *FK* into the *Knowledge* dimension. The red dashed vector indicates the position of a user within the SFB. The green continuous arrow indicates the position of a user outside the SFB.

Given that *PR* and *FK* cannot be ascertained beforehand but only in hindsight, the rule-based policy focuses on maximizing the *RUE* and *TK* dimension, 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 such that the user understands the system's choice.

Following an initialization phase (first five argu-

[2]Please note that this illustration serves solely explanatory purposes, and thus is reduced to a three dimensional space (by merging *TK* and *FK* and that the actual form and structure of the SFB may deviate.

ment 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 SFB-Model and Policy Integration into the ADS

In the following, the relevant components of the ADS, namely the knowledge base and dialogue model, focusing on the exemplary integration of our SFB-Model. In order to combine the presented model with existing argument mining approaches, ensuring its adaptability with respect to discussed topics, we adhere to the bipolar argument annotation scheme introduced Stab and Gurevych (2014)[4]. This scheme encompasses argument components (nodes), structured in the form of bipolar argumentation trees. The overall topic represents the root node in the graph. We consider two relationships between these nodes: *support* or *attack*. Each component, excluding the root node (which has no relation), has exactly one unique relation to another component. This results in a non-cyclic tree structure, wherein each node, or "parent", is supported or attacked by its "children". If no children exist, the node is a leaf and marks the end of a branch.

Move	Description	SFB Dim
why_{pro}	Request pro argument	r_k, tk_k
why_{con}	Request con argument	r_k, tk_k
$suggest$	Suggest any argument	r_k, tk_k
$prefer$	Prefer current argument	r_k
$reject$	Reject current argument	r_k
$know$	Current argument is already known	$tk_{k,i}$[3]
$false$	Current argument is incorrect	fk_k
$exit$	Terminates the conversation	

Table 1: Description of potential user actions along with their corresponding impact on SFB dimensions.

Furthermore, the SFB-Model necessitates semantically clustered arguments, wherein each argument pertains to one or more clusters related to the discussed topic. Given that an argument can encompass multiple aspects of a topic, it may belong to several overlapping clusters (Daxenberger et al., 2020). Every argument directly addresses one or more clusters. Since each argument component targets the preceding parent, it indirectly refers to all preceding parents. Consequently, we stipulate that each argument component inherits the clusters of its preceding nodes, meaning it indirectly encompasses all clusters that its parent addresses, whether directly or indirectly. Notably, the root node is not affiliated with a cluster.

In this ADS, a sample debate on the topic *Marriage is an outdated institution* provides a suitable manually clustered argument structure. It serves as the knowledge base for the arguments and is sourced from the *Debatabase* of the idebate.org[5] website. It consists of a total of 72 argument components, their corresponding relations, and is encoded in an OWL ontology (Bechhofer, 2009) for further use. In each $whypro/con$ move, a single supporting/attacking argument component is presented to the user. To prevent the user from being overwhelmed by the amount of information, the available arguments are presented to the users incrementally upon their request. In order to integrate the SFB-Model 3.2, 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[6] possible moves (actions) the user is able to choose from. This allows the user to navigate through the argument tree and inquire

[4]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.

[5]https://idebate.org/debatabase (last accessed July 23rd, 2022). Material reproduced from www.idebate.org with the permission of the International Debating Education Association. Copyright © 2005 International Debate Education Association. All Rights Reserved.

[6]Only moves that are relevant for the SFB-Model are shown. Other moves are not listed due to their mere navigational/meta-informational purposes.

for more information. The determiners show which moves are available depending on the position of the current argument.

As shown in Table 1, r_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. 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_2 is addressed and the old cluster k_1 is no longer addressed). The user's spoken input is captured through browser-based audio recording using the Google Speech Recognition API. Subsequently, it is processed by an NLU framework (Abro et al., 2022) that employs an intent classifier based on a BERT Transformer Encoder (Devlin et al., 2019) and a bidirectional LSTM classifier. After recognizing a user move, the spoken system response is presented using speech synthesis provided by the Google Web Speech API. An exemplary dialogue is shown in Appendix A.1.

5 User Study

We conducted a user study from October 4$^{\text{th}}$ to 15$^{\text{th}}$, 2022, involving 60 participants. The participants were divided into two groups: one group was presented with arguments based on their interests (referred to as the "interest" group), whereas the other group was presented with arguments that might challenge their existing beliefs (referred to as the "SFB-breaking" group). In the interest group, the system presented arguments that precisely matched the user's requests. If a loss of interest was detected (modeled by an interest model (Aicher et al., 2022a)), the system suggested arguments that aligned best with the user's preferences and interests. This interest policy is based on our previously introduced interest model (Aicher et al., 2022a) and adapted accordingly. In the SFB-breaking group, the system presented arguments based on the system policy described in Subsection 3.3. Consequently, the arguments presented to the SFB-breaking group might have differed in polarity and/or cluster from the original user request. The primary objective of this study was to address the following research questions: 1) Can the proposed system policy effectively break a user's SFB? 2) What are the discernible differences in the overall SFB dimensions between the two participant groups? To investigate these research questions, we formulated the following hypotheses to be tested during the study:

- H1 Participants in the SFB-breaking (interest) group exhibit a lower (higher) probability of being caught in an SFB after the interaction.
- H2 The exploration behavior of the SFB-breaking group changed during the interaction.

These hypotheses were designed to assess the effectiveness of the system policy in breaking the users' SFBs and to explore potential differences in SFB dimensions between the two groups. The study was conducted in a laboratory setting at a university, involving international participants who possessed a sufficient level of proficiency in English. Including the introductory phase and the completion of pre- and post-questionnaires, the entire study duration was estimated to be one hour. Participants were compensated with a payment of 10$, which corresponded to an hourly rate of 10$/hour. After a brief introduction to the system, including a short text and instructions on how to interact with it, participants were required to answer two control questions. These questions served as a means to verify their understanding of how to interact with the system. Only participants who successfully passed this test were allowed to proceed to a test interaction with the system.

During the "real" interaction, participants were instructed to listen to at least 20 arguments[7]. Participants were not informed about the underlying SFB or Interest Model. They were only informed that the ADS might provide suggestions on its own, and they could return to the previous argument if they did not approve.

Throughout the study, the following data was collected: Self-assessment questionnaire (P.851, 2003), Calculated SFB-values: *RUE*, *PR*, *TK*, and *FK* (for each cluster k), Participants' opinions and interests regarding the topic of discussion, set of heard arguments, dialogue history. Strict adherence to data protection regulations and participant anonymity was maintained throughout the study. Participants had the freedom to withdraw from the

[7]This minimum ensured a sufficient amount of data was collected to analyze the different system policies.

study at any time. The study was approved by an Institutional Review Board (IRB) after thorough ethical review and met all internal guidelines due to the solely cooperative, non-persuasive design of the user study.

6 Results

The user study involved 60 participants, ranging in age from 22 to 41 years. The participants' average age was 28.45 (with a standard deviation (SD) of 4.11). The two participant groups each consisted of 30 individuals (SFB-breaking: 7 females, 23 males; interest: 10 females, 20 males). Both groups exhibited similar levels of experience with spoken dialogue systems, rated on a 5-point Likert scale where 1 represented "No experience" and 5 represented "Very much experience": interest group at 2.40 (SD 0.89); SFB-breaking group at 2.13 (SD 1.04).

On average, participants spent approximately 33.87 minutes engaged in interactions with the system (interest group: 33.99 min (SD 7.74), SFB-breaking group: 33.75 min (SD 5.96)). Throughout the interaction, participants were presented with an average of 22.02 arguments (interest group: 21.73 (SD 4.00), SFB-breaking group: 22.30 (SD 3.54)). In Table 2, we present the mean values for all dimensions of both groups across all clusters. Given the paper's limited scope, our primary focus lies on the weighted overall means for each SFB dimension, calculated by averaging across all clusters (subtopics). Exemplary clusterwise results are provided in Appendix A.2. Notably, the SFB-breaking group displayed significantly larger values for all dimensions: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)*, and *False Knowledge (FK)* when compared to the interest group.

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p_{corr} value	r
RUE	0.30	0.28	**0.47**	0.26	<0.001	0.92
PR	0.78	0.20	**0.80**	0.19	<0.001	0.45
TK	0.28	0.18	**0.31**	0.25	<0.001	0.61
FK	0.97	0.09	**0.99**	0.05	0.008	0.39

Table 2: Means and SD of all SFB dimensions over all cluster for for both groups. Bold values indicate statically significant differences with respective Bonferroni corrected p_{corr} values and effect sizes r.

To ascertain the statistical significance of these findings, we employed the non-parametric Mann-Whitney U-test for two independent samples (McKnight and Najab, 2010). This choice was made due to the deviation of group means from normal distribution, as indicated by the Shapiro-Wilk test. Given that we are considering four dimensions, we applied the Bonferroni correction to account for multiple comparisons, thereby adjusting the p-value (represented as p_{corr}). The most substantial and statistically significant distinction was observed in the dimension of *Reflective User Engagement (RUE)* ($p_{corr} < 0.001$), as indicated by a very high effect size of 0.92 ($0.5 < r < 1$). Similarly, a significant difference ($p_{corr} < 0.001$) with a high effect size ($0.5 < r = 0.61 < 1$) was noted for *True Knowledge (TK)*. Concerning *Personal Relevance (PR)* and *False Knowledge (FK)*, the differences were also found to be highly significant, exhibiting a medium effect size ($0.3 < r < 0.5$).

Regarding the "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"), the difference between the two groups is insignificant (interest: 3.67 [SD 0.71], SFB-breaking: 3.47 [SD 0.82]; $p_{corr} = 0.986$). Similarly, the difference in their "pre-opinion" (rated on a scale of 1 to 5, where 1 represented "Totally disagree" and 5 represented "Totally agree") is also insignificant (interest: 3.09 [SD 0.93]; SFB-breaking: 2.78 [SD 0.83]; $p_{corr} = 0.308$). During the interaction, approximately 36.67% (11 out of 30) participants changed their opinion (from pro to con or vice versa) in the SFB-breaking group, compared to 6.67% (2 out of 30) in the interest group. Regarding the "post-interest" (measured after the interaction), a significant difference with $p_{corr} = 0.024 < 0.05 = \alpha$ is notable (interest: 3.20 [SD 1.16], SFB-breaking: 3.97 [SD 0.89]). Similarly, the "post-opinion" also exhibits a significant difference (interest: 3.63 [SD 0.96], SFB-breaking: 3.07 [SD 0.87], $p_{corr} = 0.048$, $r = 0.29$).

To determine the significance of the difference between pre- and post-measurements, we utilized the non-parametric Wilcoxon signed rank test (Woolson, 2007) for paired samples and Bonferroni corrected p-values p_{corr} based on a set of four comparisons. For the SFB-breaking group, both interest and opinion showed significant differences before and after the interaction (interest: $p_{corr} = 0.006$, $r = 0.38$; opinion: $p_{corr} = 0.036$, $r = 0.30$). In the interest group, the pre- and

post-interest also exhibited significant differences ($p_{corr} = 0.006, r = 0.39$).

Considering the user moves, a significant difference between both groups becomes evident. In the interest group, a pro (con) argument was requested 297 (172) times. Only in 15% of all argument requests, interest group users asked for an argument which did not align with their own opinion. In the SFB-breaking group, a con (pro) argument was requested 117 (90) times. Furthermore, in 71 (82) instances, the ADS opted to present a con (pro) argument. Particularly towards the end, the SFB-breaking group tended to request arguments without specifying polarity, and if polarity was specified, it contradicted the user's opinion in 43% of all requests. In the interest group, arguments were rarely rejected (3) and mostly preferred (87). In the SFB-breaking group, suggested arguments were rejected 65 times and explicitly preferred 71 times. Moreover, participants in the SFB-breaking (interest) group requested to return to the previous argument in only 8 (1) cases.

7 Discussion

In the following the results of our study (Section 6), particularly regarding our two hypotheses (refer to Section 5) are discussed.

7.1 Validation of Effectiveness of SFB-breaking policy (H1):

The significant differences in all overall dimensions between both groups can be attributed to the substantial disparity in polarity and the corresponding clusters to which the heard arguments belonged, despite the nearly similar number of heard arguments. While the interest group was exclusively exploring arguments of the requested polarity and the estimated most interesting clusters, the SFB-breaking group encountered arguments strategically chosen to break the SFB of the user. Consequently, participants in the interest group primarily requested arguments aligning with their pre-existing opinions. In contrast, the SFB-breaking group encountered arguments of both polarities, elucidating the significant difference in the overall *Reflective User Engagement (RUE)*. These observations further validate the hypothesis that users tend to remain within their SFBs while exploring contentious topics unless proactively motivated to consider opposing viewpoints. The substantial difference in *True Knowledge (TK)* across all clusters is a result of the SFB-breaking system's tailored policy, which aims to present arguments spanning as many clusters as possible to encompass diverse facets of the topic. In contrast, the interest policy concentrates on clusters aligned with the user's interest, offering arguments accordingly.

Significant variations in *Personal Relevance (PR)* are also evident, even accounting for differences between individual clusters, notably contingent on the number of arguments heard from each cluster. Participants who explored a greater number of clusters in a balanced manner tended to exhibit notably higher *Personal Relevance (PR)* on average. Similarly, disparities are discernible among the individual clusters concerning *False Knowledge (FK)*. Out of the nine instances of $false$ moves, merely two were initiated by participants in the SFB-breaking group. Hence, aligning with our hypothesis, the outcomes affirm that participants in the SFB-breaking (interest) group demonstrated a notably lower (higher) likelihood of being caught in an SFB after the interaction.

7.2 Change of exploration behaviour (H2):

In the initial stage of the interaction, the first five arguments presented by the ADS were selected solely based on the user's requests. During this phase, both groups exhibited a tendency to seek arguments that aligned with their pre-existing opinions. However, a shift in behavior was observed among the SFB-breaking group participants after being repeatedly exposed to arguments of opposing polarity. On average, after the eleventh argument, SFB-breaking users began to request pro and con arguments almost equally or no longer specified the polarity. Interestingly, with the exception of one case, participants from the SFB-breaking group continued the interaction and did not revert to the previous argument. This suggests that the participants appeared to be more motivated by the system's suggestions to explore differing viewpoints and facets. This observation is further supported by the heightened *Personal Relevance (PR)* of the corresponding clusters. Conversely, participants in the interest group returned to the previous argument when they did not perceive the corresponding cluster as personally relevant.

Within the SFB-breaking group, participants expressed a preference for and rejection of the proposed arguments almost equally, with approximately a third changing their opinion, resulting

in a relatively neutral post-opinion. In contrast, the interest group predominantly indicated their preference for arguments and rarely rejected any. The reinforcement of their pre-existing opinions becomes particularly evident as the interest group encountered over twice as many pro arguments as con arguments, and only two participants altered their stance on the topic. The comparatively diminished level of interest after the interaction in the interest group could potentially be attributed to a saturation effect. Conversely, the SFB-breaking group exhibited an elevated post-interest, indicative of heightened engagement and a greater willingness to explore additional aspects.

In conclusion, it is evident that the exploration behavior exhibited by the SFB-breaking group demonstrates a significant improvement in balance concerning clusters and polarity. To sum up, our findings corroborate our initial hypotheses and demonstrate that our SFB-breaking policy takes us closer to achieving our goal of assisting users in critically evaluating information on a contentious topic.

8 Limitations

However, this work has certain limitations that could be addressed in future research. First, the sample size of our study is relatively small, potentially affecting the generalizability of our findings. In future endeavors, a study (e.g., through crowdsourcing) with a larger sample size could yield more robust data, enabling us to refine the SFB margins and enhance the validity of our approach. Second, given that the SFB-Model is a novel concept, it is presently constrained to four dimensions. Subsequent research could explore additional dimensions that may prove pertinent in various scenarios and applications. Additionally, finding ways to implicitly estimate both PR and TK, which can only be determined retrospectively, would be advantageous. This could involve leveraging common sense knowledge bases and employing fake news detection techniques. Third, while our study demonstrates the proof-of-principle for the effectiveness of a rule-based policy to break SFB, it is limited to static, predefined rules, rendering it relatively inflexible. In future work, we intend to delve into more advanced machine learning techniques, such as reinforcement learning. This would enable us to personalize and adapt these strategies based on the user's verbal and non-verbal feedback, thereby ensuring the user's satisfaction and sustaining their willingness to engage in the dialogue.

9 Conclusion and Future Work

In this work, to the best of our knowledge, we introduce a novel approach to break the user's SFB. After shortly explaining the underlying SFB-Model, we define a rule-based system policy to break the respective user SFB during a cooperative dialogue with an argumentative dialogue system and validate it in a laboratory user study. The study results strongly indicate the effectiveness of the proposed system policy in reducing the likelihood of being stuck in an SFB compared to a policy that prioritizes the users' greatest interest. Moreover, the study revealed significant changes in users' exploration behaviors during the interaction. In particular, the SFB-breaking participants requested arguments of both polarities almost equally often after the ADS pointed out that the previous exploration seemed to be one-sided. These findings emphasize the influence of the system policy on users' exploration behaviors and opinions, further highlighting the success of the proposed approach in mitigating SFB tendencies and fostering open-mindedness in an argumentative dialogue. In future research, we will augment our system's policy by incorporating sophisticated techniques for perceiving and interpreting the user's non-verbal social signals (gestures, facial expressions) in real-time during the interaction. Building upon estimation methods for sentiment and emotion recognition, we aim to leverage Reinforcement Learning to optimize the system's policy, enabling it to dynamically adapt to each individual user's motivation and effectively engaging the users to recognize and overcome their SFB.

In conclusion, this paper highlights the importance of addressing SFBs in argumentative dialogues and takes us a step closer to enabling users to build a well-founded opinion and foster critical, reflective thinking, and open-mindedness in their interaction with cooperative ADS.

Acknowledgements

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

References

Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems*, 242:108318.

Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker, and Stefan Ultes. 2022a. User interest modelling in argumentative dialogue systems. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 127–136, Marseille, France.

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2021a. Determination of reflective user engagement in argumentative dialogue systems.

Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2022b. Towards modelling self-imposed filter bubbles in argumentative dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4126–4134, Marseille, France. European Language Resources Association.

Annalena Aicher, Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2021b. Opinion building based on the argumentative dialogue system BEA. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pages 307–318. Springer.

Annalena Bea Aicher, Daniel Kornmüller, Wolfgang Minker, and Stefan Ultes. 2023. Self-imposed filter bubble model for argumentative dialogues. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, pages 1–11.

Armen E Allahverdyan and Aram Galstyan. 2014. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.

Bharat N Anand. 2021. The us media's problems are much bigger than fake news and filter bubbles. *Domestic Extremism*, page 138.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Sean Bechhofer. 2009. Owl: Web ontology language. In *Encyclopedia of Database Systems*, pages 2008–2009. Springer.

Lisa Chalaguine and Anthony Hunter. 2021. Addressing popular concerns regarding covid-19 vaccination with natural language argumentation dialogues. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 59–73, Cham.

Lisa A. Chalaguine and A. Hunter. 2020. A persuasive chatbot using a crowd-sourced argument graph and concerns. In *COMMA*.

Johannes Daxenberger, Benjamin Schiller, Chris Stahlhut, Erik Kaiser, and Iryna Gurevych. 2020. Argumentext: argument classification and clustering in a generalized search scenario. *Datenbank-Spektrum*, 20(2):115–121.

Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Sci Rep*, 7(40391):1–9.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tim Donkers and Jürgen Ziegler. 2021. The dual echo chamber: Modeling social media polarization for interventional recommending. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 12–22, New York, NY, USA. Association for Computing Machinery.

Axel G. Ekström, Diederick C. Niehorster, and Erik J. Olsson. 2022. Self-imposed filter bubbles: Selective attention and exposure in online search. *Computers in Human Behavior Reports*, 7:100226.

Hans Gelter. 2003. Why is reflective thinking uncommon. *Reflective Practice*, 4(3):337–344.

Emmanuel Hadoux, Anthony Hunter, and Sylwia Polberg. 2022. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. *Argument & Computation*, 14:1–53.

Hsieh-Hong Huang, Jack Shih-Chieh Hsu, and Cheng-Yuan Ku. 2012. Understanding the role of computer-mediated counter-argument in countering confirmation bias. *Decision Support Systems*, 53(3):438–447.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Terry Lee. 2019. The global rise of "fake news" and the threat to democratic elections in the usa. *Public Administration and Policy*, 22(1).

Erin A Maloney and Fraulein Retanal. 2020. Higher math anxious people have a lower need for cognition and are less reflective in their thinking. *Acta psychologica*, 202:102939.

Mark Mason. 2007. Critical thinking and learning. *Educational philosophy and theory*, 39(4):339–349.

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

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220.

ITU-T Recommendation P.851. 2003. Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003). International Telecommunication Union.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Richard W Paul. 1990. Critical and reflective thinking: A philosophical perspective. *Dimensions of thinking and cognitive instruction*, pages 445–494. Publisher: North Central Regional USA.

Richard E Petty, Pablo Briñol, and Joseph R Priester. 2009. Mass media attitude change: Implications of the elaboration likelihood model of persuasion. In *Media effects*, pages 141–180. Routledge.

Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on facebook. *Available at SSRN 2795110*.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents - 8th International Workshop on Spoken Dialog Systems*, pages 45–52.

Ariel Rosenfeld and Sarit Kraus. 2016. Strategical argumentative agent for human persuasion. In *ECAI'16*, pages 320–328.

Christina Schwind and Jürgen Buder. 2012. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior*, 28(6):2280–2290.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, and Lilach Edelstein. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510.

Constanza Villarroel, Mark Felton, and Merce Garcia-Mila. 2016. Arguing against confirmation bias: The effect of argumentative discourse goals on the use of disconfirming evidence in written argument. *International Journal of Educational Research*, 79:167–179.

Abdul Waheed, Muskan Goyal, Nimisha Mittal, Deepak Gupta, Ashish Khanna, and Moolchand Sharma. 2021. BloomNet: A robust transformer based model for bloom's learning outcome classification. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 209–218, Trento, Italy. Association for Computational Linguistics.

RF Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials*, pages 1–3.

A Appendix

A.1 Example Interaction

In Table 3, we present a segment of an exemplary dialogue with the argumentative dialogue system, following the SFB-breaking policy. It illustrates an artificial interaction between the system and the user regarding the topic *Marriage is an outdated institution*, sourced from the *Debatabase* on the idebate.org[8] website.

We assume that the interaction has been ongoing for a substantial duration, and the user's previous interactions suggest a likelihood of being trapped in an SFB due to consistently requesting attacking arguments.

Throughout the illustrated interaction, the user's requests for con/pro arguments impact two SFB dimensions: the *Reflective User Engagement* (r_k) and the *True Knowledge* (tk_k), within the corresponding SFB-vector $\overrightarrow{sfb_k}$. Additionally, the value for *False Knowledge* (fk_k) is updated due to the user's contradictory knowledge statement.

A.2 Exemplary clusterwise Results

Due to the limited scope of the paper, we focused on the overall SFB dimension results. Nonetheless, as we observed variations among individual clusters, we present the results of two specific clusters below.

Table 4 displays the mean values for each SFB dimension in both groups for two particular clusters (subtopics) of the discussed theme "Marriage is an outdated institution." Cluster 4 ("Children") comprises 22 arguments (one of the largest clusters), while Cluster 7 ("Law") includes only 6 arguments, making it one of the smallest clusters.

Similar to the overall cluster means (refer to Table 2), we observe significant differences with

[8] https://idebate.org/debatabase (last accessed 23rd 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.

Actor	Utterance	Move	SFB Dimension
System	...		
User	*Could you please provide an attacking argument?*	why_con	r_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_con	r_k
System	*On the pro side, it's 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.*		

Table 3: Artificial dialogue between the ADS and the user on the topic *"Marriage is an outdated institution"*. The table includes actor, utterance, respective moves, and updated SFB dimensions of cluster k.

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p_{corr} value	r
r_4	0.35	0.20	**0.66**	0.12	<0.001	0.63
pr_4	0.71	0.21	**0.80**	0.12	0.007	0.23
tk_4	0.25	0.14	**0.47**	0.64	<0.001	0.73
fk_4	0.93	0.13	**1.00**	0.00	0.044	0.21
r_7	0.21	0.39	**0.82**	0.19	<0.001	0.73
pr_7	0.74	0.15	**0.81**	0.18	<0.001	0.35
tk_7	0.45	0.26	**0.83**	0.15	<0.001	0.77
fk_7	0.92	0.14	**1.00**	0.00	0.021	0.32

Table 4: Means and SDs of all SFB dimensions for two clusters (4 = "Children", 7 = "Law") for both groups. Bold values indicate statically significant differences with respective Bonferroni corrected p values and effect sizes r.

small to high effect sizes in each dimension. Furthermore, noticeable differences are evident between individual clusters, as illustrated in Table 4. Particularly concerning smaller clusters, we discern that our SFB-breaking policy has a moderate to large effect on each dimension. This can be attributed to the fact that our SFB-breaking policy aims to explore all clusters in a balanced manner, whereas the interest policy only targets clusters of user interest.