Towards Breaking the Self-imposed Filter Bubble in Argumentative Dialogues

Anonymous ACL submission

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 bubble" (SFB) are a great challenge in cooperative argumentative dialogues with the goal 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' SFBs within the course of the interaction. This strategy is implemented in an argumentative dialogue system and evaluated in a laboratory user study with 60 participants to show its validity and applicability. The findings of the study 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 to easily access requested information from online sources such as search engines or social media platforms. Especially with regard to more complex interactions two important phenomena can be observed that can result in an 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, 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 behaviour leads to so-called "Self-imposed Filter Bubbles" (SFB) (Ekström et al., 2022; Aicher et al., 2022c) and "echo chambers" (Quattrociocchi et al., 2016; Anand, 2021; Donkers and Ziegler, 2021) which are both 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 a recently presented 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 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 give a short overview of existing literature on the main aspects of the herein presented work, *Confirmation Bias and Self-imposed Filter Bubbles* and *Argumentative Dialogue Systems*.

2.1 Confirmation Bias and Self-imposed Filter Bubbles

As previously pointed out, the users' seeking or interpreting of evidence in ways that are partial to their existing beliefs, expectations, or a hypothesis in hand is called confirmation bias (Nickerson,

1998). Allahverdyan and Galstyan (2014) describe confirmation bias as the tendency to acquire or evaluate new information in a way that is consistent with one's preexisting beliefs. To resolve the confirmation bias of a user in the context of decision making processes Huang et al. (2012) propose the usage of computer-mediated counter-argument. Schwind and Buder (2012) regard preference-inconsistent recommendations as a promising approach to trigger critical thinking. Still, if too many counter-arguments are introduced this could lead to unwanted effects negative emotional consequences (annoyance, confusion) (Huang et al., 2012). According to Paul (1990) if users think critically in a *weak sense*, this implies reflecting about positions that are different from the one's own (Mason, 2007), but tending to defend the own view without reflection (Paul, 1990). Critical thinking in a *strong sense* means to reflect one's own opinion as well. The energy and effort (Gelter, 2003) required for this strong critical reflection is often not present due to a lack of people's *need for cognition* (Maloney and Retanal, 2020). Due to the users' tendency to defend their own view (Paul, 1990), a system which confronts them with an opposing stance might not lead to critical reflection but rather the opposite. Consequently, Huang et al. (2012) stress the need for an intelligent system which is able to adapt the frequency, timing and choice of the counter-arguments. To provide such a system, it is crucial to develop a model, which can be adapted to the user. An exemplary approach for such a model is introduced by Del Vicario et al. (2017), who study online social debates and try to model the related polarization dynamics based on confirmation bias mathematically. In contrast, we aim to model the cause of this bias, the so-called "Self-imposed Filter Bubble" (SFB) (Ekström et al., 2022). Therefore, we build upon the work of (Aicher et al., 2022c) and use their proposed indicators to model the user's SFB. To the best of our knowledge, we are the first to define a system policy which aims to help the users to overcome their SFBs in a cooperative argumentative dialogue.

2.2 Argumentative Dialogue Systems

According to Villarroel et al. (2016) 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. Thus, the respective ADS in which the SFB-model is incorporated into, should 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 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 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 and Hunter (2021). 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). In contrast, the system of Aicher et al. (2021b) is based upon a cooperative exploration of arguments and offers the users the possibility to state their preferences and thus, offers a more suitable basis than formerly described ADS. Its extension with a corresponding NLU (Aicher et al., 2022a) and speech interface of this system serves as a basis for the herein utilized spoken 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. (2022c) which is founded on a well-established framework in persuasion research, the likelihood model (ELM) (Petty et al., 2009). It consists of four dimensions, which span a four-dimensional space: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*. The **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 **PR** refers to the user's individual assessment of how relevant a subtopic is with regard to the topic of the discussion. The bigger the PR of a certain cluster is, the higher is the user's motivation to explore arguments belonging to it. The **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. The user should to explore as much information as possible, as this 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 **FK** describes the incorrect information a user has on a certain topic[1]. If the user is misinformed on certain aspects, it increases the probability of being stuck in an SFB and reluctant towards contradicting information and viewpoints. Thus the bigger the inverted FK, the smaller is the chance to be trapped in an SFB.

3.2 SFB-Model

Argumentative discussions are complex and consist of a lot of different subtopics, which contain arguments referring to the same content-related aspects. For each of these so-called "clusters" corresponding SFB vectors $\overrightarrow{sfb_k} = (pr_k, r_k, tk_k, fk_k)^T,\ k \in \mathbb{N}$ is defined, which finally make up the overall SFB vector $\overrightarrow{SFB_k}$ of the whole discussion topic. It is crucial to distinguish between the SFB and SFB-vector of a user (see Figure 1. The SFB-vector is defined as a vector that has its origin in the origin of the coordinate system and whose end is the position of the user in the four-dimensional space at the current state of the interaction. The SFB is the area in the four-dimensional space that represent a certain probability of users to be in an SFB. In Figure 1 an exemplary sketch of this vector and the respective SFB of one cluster are shown. As a four-dimensional vector cannot be displayed, it was split for a better illustration in two different z_1-components tk_k and fk_k. Please note that this sketch is for illustrative purposes only and the "real" shape and structure of the SFB marked in light blue may differ. Especially, as it is hard to define distinct margins, we model a probability for a user to be inside or outside the SFB. The smaller the SFB-vector, the higher the probability that the user is inside the SFB. The longer the SFB vector and the more it extends beyond the SFB, the lower the probability that the user is within the SFB. The over-

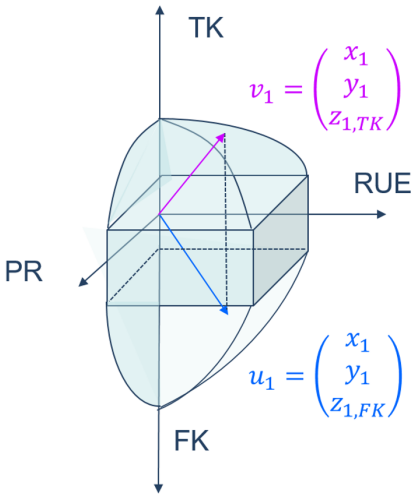

Figure 1: Schematic sketch of a clusterwise SFB-vector and SFB for a cluster k. The box indicates, the probability of a filter bubble is very high near the origin and if a dimension is close to zero. For better illustration the four-dimensional SFB-vector is displayed in two split components which only differ regarding their z_1 component. Whereas the blue vector displays the tk_k in the z_1-component, the violet one displays the fk_k. The x_1 component depicts the reflective user engagement r_k and y_1 the personal relevance pk_k. The blue areas denote the SFB.

all SFB vector $\overrightarrow{SFB} = (PR, RUE, TK, FK)^T$, consists of the overall cluster values for each dimension, derived from a weighted mean calculation (BLIND, 2023)[2].

3.3 SFB-breaking policy

Building upon the model described in Subsection 3.2 we propose a rule-based system policy aiming to break the user's SFB. Based on the data of a previous crowd-sourcing user study, we determined how the SFB dimensions changed with two different system policies. The first one was the same interest policy presented in Section 5, which chooses

[1]Without loss of generality, the information in the system's database is defined as factually correct and consequently, information contradicting the former to be incorrect.

[2]Referenced literature is blinded due to double blind review requirements.

arguments according to the estimated user's interest. The second one randomly presented arguments from the remaining arguments. The calculated averages across all participants serve as reference points to define areas with a higher probability to being stuck in an SFB (very high probability = interest average; medium probability = random average). Since PR and FK cannot be determined in advance, but only retrospectively, the rule-based policy aims to maximize the RUE and TK dimensions, which can be calculated in advance. In case the PR or TK change after a suggested argument has been presented, the related arguments regarding polarity and corresponding clusters are prioritized. In order to ensure logical consistency, potential argument candidates are selected have to have a relation to the requested argument by either being siblings or by having the greatest possible overlap in belonging to the same clusters. Identified candidates are then compared to the user-selected argument in the corresponding RUE and TK dimensions and finally the argument with the maximum values is presented. In case the system chooses an argument differing from the user choice, the corresponding system response includes an explanation why the system chose a different argument. After an initialization phase (first five argument requests) to detect and reward changes in users' exploration behaviors, after each interaction turn the current user SFB-vector is compared to the data-based SFB-margins (interest, random). If the SFB-vector remains within the first area (below interest margin) the ADS will opt for selecting the best available argument in each turn. If the SFB-vector is within the second region (below random margin), a decision is made based on the recent changes in the SFB vector over the past three interaction turns, determining whether the system should provide a suggestion or presents the requested argument. If the SFB-vector exceeded the random margin, the ADS presented the requested argument under the condition 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 (knowledge base and dialogue model) of the ADS with regard to the exemplary integration of our model are introduced. To be able to combine the presented model with existing argument mining approaches to ensure its flexibility in view of discussed topics, we follow the bipolar argument annotation scheme introduced by Stab and Gurevych (2014)[4]. It consists of argument components (nodes), which are structured in the form of bipolar argumentation trees. The overall topic of the debate represents the root node in the graph. We consider two relations between these nodes: *support* or *attack*. Each component apart from the root node (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" is supported or attacked by its "children". If no children exist, the node is a leaf and marks the end of a branch. Furthermore, the SFB-Model requires semantically clustered arguments, such that each argument belongs to one or more clusters of the discussed topic. As an argument can address more than one aspect of a topic, it may belong to multiple overlapping clusters (Daxenberger et al., 2020). Each argument directly addresses one or more clusters. As each argument component targets the predecessor above it, it refers indirectly to all predecessing parents. Therefore, we define that each argument component inherits the clusters of its preceding nodes, i.e. it indirectly addresses all clusters its parent directly or indirectly addresses. The root node does not belong to a cluster. In this ADS a sample debate on the topic *Marriage is an outdated institution* provides a suiting manually clustered argument structure. It serves as knowledge base for the arguments and is taken 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 on 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 in

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

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

Move	Description	Determiners	SFB Dim
why_{pro}	Request for a pro argument	If supporting argument exists	r_k, tk_k
why_{con}	Request for a con argument	If attacking argument exists	r_k, tk_k
$suggest$	Suggest another argument (no polarity)	If unheard arguments exist	r_k, tk_k
$prefer$	State agreement/preference for current argument	Always	r_k
$reject$	State disagreement/rejection of current argument	Always	r_k
$know$	States that current argument is already known	Always	$tk_{k,i}$[3]
$false$	States that current argument is incorrect	Always	fk_k
$exit$	Terminates the conversation	Always	

Table 1: Description of the possible user moves with corresponding determiners and influenced SFB dimension.

turns, consisting of a user action and corresponding natural language answer of the system. The system response is based on the original textual representation of the argument components, which is embedded in moderating utterances. Table 1 shows the required[6] possible moves (actions) the user is able to choose from. Thereby the user is able to navigate through the argument tree and enquire 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. This does not apply for PR, which does not refer directly to the dialogue content but rather displays a meta reflection. As pr_k is not directly referring to the argument but the respective cluster this information is requested in a separate pop-up window during the interaction. In order not to annoy the user (as the cluster might be the same over a certain number of moves), we update pr_k whenever the corresponding clusters change (new cluster k_2 is addressed, old cluster k_1 is not addressed anymore). The user's spoken input is captured with a browser-based audio recording using the Google Speech Recognition API. Afterwards it is processed by an NLU framework (Abro et al., 2022) using an intent classifier based on a BERT Transformer Encoder (Devlin et al., 2018) and a bidirectional LSTM classifier. After a user move is recognized, the spoken system response is presented using the speech synthesis provided by Google Web Speech API. An exemplary dialogue is shown in the Appendix A.1.

5 User Study

We conducted a user study from October 4th to 15th, 2022, involving 60 participants. The participants were divided into two groups: one group was presented 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 (modelled by an interest model (Aicher et al., 2022b)), the system suggested arguments that aligned with the user's preferences and interests best, taking into account the user's preference or rejection of previous arguments. This interest-policy is based on the findings of (Aicher et al., 2022b) and adapted accordingly to our ADS. 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 will exhibit a lower (higher) probability of being entrapped in 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 user SFBs and to explore potential differences in SFB dimensions between the two participant groups.

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

The study was conducted in a laboratory setting at a university, involving international participants who possessed a sufficient English proficiency level. Including the introductory phase and the completion of pre- and post-questionnaires, the whole study duration was estimated to be an 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 pass 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 study at any time. The study was approved by an Institutional Review Board (IRB) after a thorough (ethical) review and met all internal guidelines due to the solely cooperative, non-persuasive design of the user study. The user study involved 60 participants, with ages ranging from 22 to 41 years. The average age of the participants was 28.45 (standard deviation (SD) 4.11). The two participant groups consisted of 30 individuals (SFB-breaking: 7 females, 23 males: interest: 10 females, 20 males). Both groups exhibited similar levels of experience (5-point Likert scale, where 1 represented "No experience" and 5 represented "Very much experience") with spoken dialogue systems: interest 2.40 (SD 0.89); SFB-breaking: 2.13 (SD 1.04).

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

6 Results

On average, participants spent 33.87 min in the interaction with the system (interest: 33.99 min (SD 7.74), SFB-breaking: 33.75 min (SD 5.96)) and heard 22.02 arguments (interest: 21.73 (SD 4.00), SFB-breaking: 22.30 (SD 3.54)). In Table 2 the mean values for all dimensions for both groups for all clusters are shown. Due to the limited scope of this paper, we mainly focus on the weighted overall means for each SFB dimension averaged over all clusters and describe exemplary cluster-wise results in the Appendix A.2. Remarkably, the SFB-breaking group exhibited significantly larger values for all dimensions RUE, PR TK and FK (inverted!) compared to the interest group. To determine the statistical significance we employed the non-parametric Mann-Whitney U-test for two independent samples (McKnight and Najab, 2010), as the group means were found to deviate from normal distribution according to the Shapiro-Wilk test. The most substantial and statistically significant distinction was identified in the dimension RUE ($p < 0.001$), indicated by a very high Pearson's correlation coefficient (effect size measure) of 0.92 ($0.5 < r < 1$). Furthermore, with regard to TK, a very significant difference ($p < 0.001$) was noticeable, accompanied by a high effect size ($0.5 < r = 0.61 < 1$). Regarding PR and FK the differences were also very significant with a medium effect size ($0.3 < r < 0.5$). Regarding the "pre-interest"[8] of the participants the difference between the two groups is insignificant (interest: 3.67 (SD 0.71), SFB-breaking: 3.47 (SD 0.82); p=0.493). Likewise, the difference in their "pre-opinion"[9] is insignificant (interest: 3.09 (SD 0.93); SFB-breaking:

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p 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	0.002	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 p values and effect sizes r.

[8]5-point Likert scale before the interaction, where 1 represented "No at all interested" and 5 represented "Very much interested"

[9]1 represented "Totally disagree" and 5 represented "Totally agree"

2.78 (SD 0.83) ; p=0.154). During the interaction about 36.67% (11 of 30) participants changed their opinion (from pro to con or vice versa) in the SFB-breaking group and 6.67% (2 of 30) in the interest group. Regarding the "post-interest" (after the interaction) a significant difference with $p = 0.012 < 0.05 = \alpha$ is noticeable (interest: 3.20 (SD 1.16), SFB-breaking: 3.97 (SD 0.89)). Likewise, the "post-opinion" differs significantly (interest: 3.63 (SD 0.96), SFB-breaking: 3.07 (SD 0.87), p=0.025, r=0.29). To determine whether the difference between pre/post is significant, we used the non-parametric Wilcoxon signed rank test (Woolson, 2007) for paired samples. Regarding the SFB-breaking group both interest and opinion differed significantly before and after the interaction (interest: p=0.003, r=0.38; opinion: p=0.018, r=0.30). In the interest group the pre- and post-interest differed significantly (p= $0.003, r = 0.39$). Considering the user moves a significant difference between both groups is perceivable. In the interest group 297 (172) times a pro (con) argument was requested. Only in 15% of all argument request, the argument polarity was not in-line with the participant's opinion. In the SFB-breaking group 117 (90) times a con (pro) argument was requested. Furthermore, in 71 (82) times the ADS decided to present a con (pro) argument. Especially towards the end, the SFB-breaking group tended to request arguments without polarity specification and if a polarity was specified, it contradicted the user opinion in 43% of all requests. In the interest group arguments were almost not rejected (3) and mostly preferred (87). In the SFB-breaking group 65 times suggested arguments were rejected and 71 times they were explicitly preferred. Moreover, only in 8 (1) cases did participants request to return to the previous argument in the SFB-breaking (interest) group.

7 Discussion

In the following section we discuss the results of our study presented in Section 6, especially with respect to our two previously defined hypotheses (see 5.

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

The significant differences in all overall dimensions between both groups can be explained by the large difference in the polarity and corresponding clusters the heard arguments belonged to, even though number of heard arguments is nearly similar. Whereas the interest group was only presented with the arguments of requested polarity and the estimated most interesting cluster(s), the SFB-breaking group was presented with the arguments which were best to break the SFB of the user. Consequently, the participants in the interest group predominantly requested arguments that aligned with their pre-existing opinions, whereas the SFB-breaking group was presented with arguments of both polarities, explaining the significant difference in the overall RUE. These observations further confirm the hypothesis that users are prone to stay within their SFBs during the exploration of controversial topics unless they are proactively motivated to explore opposing viewpoints. The significant difference in TK throughout all clusters is due to the fact, that the SFB-breaking system's policy is tailored present arguments which belong to as many clusters as possible, to cover more aspects of the topic. The interest policy on the other hand mainly focuses on the clusters the user is interested in and provided accordingly suited arguments. The differences in PR are also significant even though there are differences between individual clusters, particularly depending on the number of arguments heard from each cluster. Participants who explored a greater number of clusters in a balanced manner tended to have significantly higher PR on average. Likewise, differences among the individual clusters are perceived in FK. Out of the nine instances of $false$ moves, only two were initiated by participants of the SFB-breaking group. Thus, meeting our hypothesis, the results show that participants in the SFB-breaking (interest) group exhibited a significantly lower (higher) probability of being trapped in SFB after the interaction.

7.0.2 Change of exploration behaviour (H2):

In the initial stage of the interaction, the first five arguments presented by the ADS were chosen solely according the user's request. During this phase, both groups displayed a proclivity for seeking arguments aligning with their pre-existing opinions. However, a shift in behavior was observed among the SFB-breaking group after the participants were repeatedly suggested arguments of opposing polarity. On average after the eleventh argument, the SFB-breaking users started to request pro and con arguments almost equally often or did not longer specify the polarity. Interestingly, except for one case the SFB-breaking group participants contin-

ued the interaction and did not return to the previous argument. This indicates that the participants seem to have been more motivated by the system's suggestions to explore other viewpoints and aspects. This is furthermore underpinned by the increased PR of the corresponding clusters. On the other hand interest group participants returned to the previous argument, when they did not consider the corresponding cluster as personally relevant to them. In the SFB-breaking group, the participants preferred and rejected the proposed arguments nearly equally often and about a third changed their opinion resulting in a rather neutral post-opinion on average. Conversely, the interest group predominantly expressed their preference for arguments and hardly rejected any of them. The reinforcing of their pre-existing opinions becomes especially apparent as the interest group heard more than twice times more pro than con arguments and only two participants changed their opinion on the topic. The comparatively lower level of interest after the interaction in the interest group might be due a saturation effect. On the other hand, the SFB-breaking group displayed an increased post-interest indicating a heightened engagement and motivation to explore additional aspects. Therefore, it can be observed that the exploration behavior shown by the SFB-breaking group experiences a significant improvement in balance regarding clusters and polarity. To summarize, our findings support our initial hypotheses and show that our SFB-breaking policy brings us closer to our aim to support users in a critical scrutinizing of information on a controversial topic.

8 Limitations

However, the previously described work is subject to some limitations that could be addressed in future research. First, the sample size of our study is rather small which may affect the generalizability of our findings. In future work a study (e.g. via crowdsourcing) with a larger sample size would provide more robust data to refine the SFB margins and enhance the validity of our approach. Second, as the SFB-Model is a novel approach, it is currently limited to four dimensions. Future research could explore additional dimensions that might be relevant to different scenarios/applications. Furthermore, as PR and TK can only be determined retropersectively, approaches to implicitly estimate both would be desirable, e.g. by involving common sense knowledge bases and face news detection techniques, Third, although our study provides proof-of-principle for the effectiveness of rule-based policy to break SFB, it is still limited to static, predefined rules and rather inflexible. In future research we will explore more sophisticated machine learning techniques, such as reinforcement learning, to personalize and adapt these strategies to the user's verbal and non-verbal feedback to maintain the user's satisfaction and willingness to continue 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 ADS and validated it in a laboratory user study. The study results strongly indicate the effectiveness 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 of 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 overcoming the latter 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 the interaction with cooperative ADS.

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, Isabel Feustel, Wolfgang Minker, and Stefan Ultes. 2022a. Towards building a spoken dialogue system for argument exploration. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1234–1241, Marseille, France. European Language Resources Association.

Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker, and Stefan Ultes. 2022b. 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. 2022c. 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.

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.

BLIND. 2023. Blind. *BLIND*, X(X):X.

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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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 and Anthony Hunter. 2021. Strategic argumentation dialogues for persuasion: Framework and experiments based on modelling the beliefs and concerns of the persuadee. In *arXiv*, volume 2101.11870.

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*.

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.

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.

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 provide a part of an exemplary dialogue exemplary with the ADS which follows the SFB-breaking policy. It shows an artifical interaction between the system and the user on the topic *Marriage is an outdated institution* taken from the *Debatabase* of the idebate.org[10] website. We assume that the interaction already went on for quite a while and the previous exploration of the user shows a probability to be stuck in an SFB due to requesting solely attacking arguments. During the course of the shown interaction the user's requests for a con/pro argument influence two SFB dimensions, r_k and tk_k of the respective SFB-vector $\overrightarrow{sfb_k}$. Furthermore fk_k is updated due to the user's statement of contradicting knowledge.

A.2 Exemplary clusterwise Results

Due to the limited scope of the paper, we concentrated on the overall SFB dimension results in this paper. Still, as we perceived some differences between the individual clusters, we show the results of two individual clusters in the following. Table 4 shows the mean values for each dimensions for both groups for the "largest" and smallest clusters:"Divorce" (Cluster 3) which is the largest cluster (22 arguments) and "Harmful Relationships" (Cluster 5) which is the smallest cluster (4 arguments). Likewise to the overall cluster means, a significant difference with $p < 0.001$ and high effect size can be perceived in the RUE and TK dimension for all individual clusters, as exemplified in Table 4. In contrast to the overall cluster means for both cluster neither PR nor FK show a significant difference, which can be explained by the large differences between the individual cluster depending on the amount of arguments having been heard by both groups. The majority of clusters (6-7 out of 10) showed a significant difference in all four SFB dimensions between the two groups.

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

Actor	Utterance	Move	SFB Dimension
System	...		
User	*Could you please provide an attacking argument?*	why_con	r_k
System	*Instead of looking at an attacking argument, I think we should also consider the opposing view as well. One could argue that religion as a whole is becoming less important and, with it, marriage is becoming less important.*	why_pro_sfb	tk_k
User	*I see, but surely you can provide more evidence right?*	why_con	r_k
System	*On the pro side it is to mention that in the last 20 years, the number of people in the UK who identify as religious has declined by 20% according to the British Social Attitudes Survey of 2007*		tk_k
User	*Wow, don't you think this statistic is outdated by now. I do not think it is correct anymore.*	$false$	fk_k
System	*I understand you are sceptical about this, but if you let me I would like to explain this further.*		

Table 3: Artificial dialogue between the ADS and the user about the topic *"Marriage is an outdated institution"*. Apart from the actor and utterance, the respective moves and correspondingly updated SFB dimensions of cluster k are indicated.

	Interest		SFB-breaking			
Asp.	M	SD	M	SD	p value	r
r_3	0.35	0.20	**0.66**	0.12	<0.001	0.7§
pr_3	0.75	0.24	0.80	0.20	0.857	0.02
tk_3	0.25	0.14	**0.47**	0.64	<0.001	0.75
fk_3	0.97	0.10	1.00	0.00	0.087	0.23
r_5	0.21	0.39	**0.82**	0.19	<0.001	0.59
pr_5	0.78	0.21	0.80	0.20	0.787	0.03
tk_5	0.45	0.26	**0.83**	0.15	<0.001	0.59
fk_5	0.99	0.07	0.99	0.06	0.981	0.02

Table 4: Means and SDs of all SFB dimensions for two clusters (3 = "Divorce", 5 = "Harmful Relationships") for both groups. Bold values indicate statically significant differences with respective p values and effect sizes r.