Self-imposed Filter Bubble Model for Argumentative Dialogues

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

Daniel Kornmüller
daniel.kornmueller@uni-ulm.de
Institute for Communications Engineering, Ulm University
Ulm, BW, Germany

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

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

ABSTRACT

During their information seeking people tend to filter out all the parts of the available information that do not fit their existing beliefs or opinions. In this paper we present a model for this "Self-imposed Filter Bubble" (SFB) consisting of four dimensions. Thereby, we aim to 1) estimate the probability of the user being caught in an SFB and consequently, 2) identify suitable clues to reduce this probability in the further course of a dialogue. Using an exemplary implementation in an argumentative dialogue system, we demonstrate the validity and applicability of this model in an online user study with 102 participants. These findings serve as a basis for developing a system strategy to break the user's SFB and contribute to a sustainable and profound reflection on a topic from all viewpoints.

CCS CONCEPTS

• **Human-centered computing** → **User models**; *User studies*; *HCI theory, concepts and models.*

KEYWORDS

Confirmation Bias, Echo Chambers, User Modeling, Computational Argumentation, Cooperative Argumentative Dialogue Systems (ADS)

ACM Reference Format:
Annalena Aicher, Daniel Kornmüller, Wolfgang Minker, and Stefan Ultes. 2023. Self-imposed Filter Bubble Model for Argumentative Dialogues. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3571884.3597131

This work is licensed under a Creative Commons Attribution International 4.0 License.

CUI '23, July 19–21, 2023, Eindhoven, Netherlands
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3597131

1 INTRODUCTION

Conversational user interfaces (CUIs), such as chatbots and conversational (voice) assistants 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, according to Pariser [31] due to filter algorithms, information content is selected based on previous online behavior which leads to cultural/ideological bubbles, so-called "Filter Bubbles". In their literature review Michiels et al. [29] focus the technological filter bubble defined as a "decrease in the diversity of a user's recommendations over time, in any dimension of diversity, resulting from the choices made by different recommendation stakeholders".

On the other hand, Nickerson [30] 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) [17] and "echo chambers" [7, 16, 34] which are both manifestations of "confirmation bias", a term typically used in psychological literature [30]. These phenomena are mutually dependent and reinforcing according to Lee [25] as the self-imposed media bubble is as a result reinforced and perpetuated using algorithmic filters delivering content aligned with presumed interests based on search histories and personal associations. Moreover, Bakshy et al. [8] 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 certain topic during the interaction with an argumentative dialogue system (ADS). We introduce a model, which enables to 1) estimate the probability of a user being in an SFB during an ongoing interaction and 2) identify suitable reference points to reduce this probability in the further course of the interaction. Our SFB model consists of four dimensions: *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 [3]. The *PR* refers to the user's individual assessment of the relevance of subtopics

with regard to the topic of the discussion. *True Knowledge* is defined as the new information the user receives on a topic by talking to the system. *False Knowledge* refers to the user's incorrect information on a topic which contradicts the verified information in the system's database.

As CUIs provide a more natural and intuitive access to requested information, SFB are very likely to be perceived in the interaction between users and CUIs and unconsciously influence this interaction. Therefore the identification of SFBs is crucial to enable the CUI to respond in a manner that reduces the potential for information asymmetry and bias. Being immediately involved in the interaction the ability of CUIs to counteract such SFBs represents a valuable component in promoting a more engaging and balanced exchange of information.

As an unbiased and critical reflected opinion building process is more likely in a cooperative dialogue between the system and user, it is important not to force new information onto the user but to find a more subtle way. Thus, the SFB model enables to identify possible points of reference (the most decisive dimensions strengthening the bubble) which can be used as starting point to break the user's SFB. The remainder of the paper is as follows: in Section 2 an overview of related literature is given. Afterwards we introduce our novel SFB model, its components and requirements in Section 3. Section 4 discusses an exemplary integration of our model in an argumentative dialogue system which is evaluated in a crowdsourcing study with online users 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 [30]. Allahverdyan and Galstyan [5] describe confirmation bias as the tendency to acquire or evaluate new information in a way that is consistent with one's preexisting beliefs. Additionally, Jones and Sugden [22] showed that a positive confirmation bias, in both information acquisition and information use, is present in an experiment in which individuals choose the"information what to buy, prior to making a decision". A neurological implication of confirmation bias is shown by Kappes et al. [23] whose results demonstrate that existing judgments alter the neural representation of information strength, leaving the individual less likely to alter opinions in the face of disagreement.

To resolve the confirmation bias of a user in the context of decision making processes Huang et al. [21] propose the usage of computer-mediated counter-argument. Furthermore Schwind and Buder [38] 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) [21]. According to Paul [32] if users think critically in a *weak sense*, this implies reflecting about positions that are different from the one's own [28], but tending to defend the own view without reflection [32]. Critical thinking in a *strong sense* means to reflect one's own opinion as well. The energy and effort [19] required for this strong critical reflection is often not present due to a lack of people's *need for cognition* [27]. Due to the users' tendency to defend their own view [32], a system which confronts them with an opposing stance might not lead to critical reflection but rather the opposite. Consequently, Huang et al. [21] 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 approach for such a model is for example introduced by Del Vicario et al. [14], who study online social debates and try to model and describe 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) [17]. To the best of our knowledge, we are the first to define measurable dimensions to describe and build up a model for this phenomenon in context of a cooperative argumentative dialogue. This cooperative setting is motivated by the findings of Villarroel et al. [41] who state that 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.

2.2 Argumentative Dialogue Systems

Due to the previously motivated cooperative approach to exchange arguments the system in which our SFB-model shall be incorporated should not try to persuade or win a debate against a user unlike most approaches to human-machine argumentation. Those approaches utilize different models to structure the interaction and are embedded in a competitive scenario. For instance, Slonim et al. [39] use a classical debating setting. Their IBM Debater 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 [37] presenting a system based on weighted Bipolar Argumentation Frameworks (wBAG). Arguing chatbots such as Debbie [35] and Dave [24] 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. [20]. In the same line, [12] 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 [11]. Furthermore, Fazzinga et al. [18] illustrate an approach towards a dialogue system architecture that uses argumentative concepts to perform reasoning and provide answers consistent with the user input, which is illustrated by the example of a user requiring information about COVID-19 vaccines. In contrast, the system of Aicher et al. [4] 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.

3 SELF-IMPOSED FILTER BUBBLE MODEL

In the following section we will give an overview of each model dimension and motivate our choice. Afterwards we explain the basic model how these dimensions form the clusterwise and overall SFB-vector of the user and necessary requirements for integration in an ADS. Please note, that we do not claim that the dimensions or our model to be complete but that it is a first approach to model SFBs.

3.1 SFB-Model Dimensions

As previously mentioned we focus on four dimensions in our model, which span a four-dimensional space: *Reflective User Engagement (RUE), Personal Relevance (PR), True Knowledge (TK)* and *False Knowledge (FK)*. We motivate this choice building upon findings in well-established state-of-the-art literature. 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" we define corresponding SFB vectors $\overrightarrow{sfb_k},\ k \in \mathbb{N}$ (one for each subtopic), 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. Furthermore, the SFB shall be areas in the four-dimensional space that indicate with which probability users are located within an SFB when their SFB-vectors lie in this area depending on predefined limits.

3.1.1 Reflective User Engagement (RUE). The elaboration likelihood model (ELM) [33], a well-established framework in persuasion research, suggests that an attitude change occurs as a result of two different information processing modes – central vs. peripheral. Westerwick et al. [42] state that if users process information via the central route, they engage carefully and thoroughly with the information, reflect on it, connect it with preexisting cognitions, and integrate it into their overall cognitive network which is what we aim for. But when lacking the motivation and ability for such effortful consideration, recipients may engage in peripheral processing and thus, not scrutinize the message content much. Therefore, the peripheral mode increases the probability for users to get stuck in their SFB. The reflective user engagement (*RUE*) describes the critical-thinking and open-mindedness demonstrated by the user when exploring a controversial topic [3]. Both, critical-thinking and open-mindedness appear as frequently suggested starting points to counteract various types of biases [6, 26, 38]. Recalling the definition of confirmation bias and SFBs, which implies the opposite of an reflective engagement, it follows that the *RUE* is very likely to have a big influence on the user's SFB-vector. Thus, the bigger the *RUE* concerning a certain cluster, the lower the probability is that the user is caught in a cluster SFB.

Building upon the approach of Aicher et al. [3] to determine the RUE, we propose a RUE calculation which takes into account the polarity and number of arguments (belonging to a cluster) a user has heard. Recalling that the RUE is defined as the user's interest in scrutinizing arguments and exploring diverging views, this can be mapped on two actions of the user by asking for more information, either on the pro or con side of the topic of the discussion. Therefore, our model requires the user to know how many arguments are available which is displayed in the graphical user interface in form of a corresponding visualization throughout the interaction. Thereby, we can deduce that unheard arguments are left out intentionally and not by mistake. Consequently, the more arguments of both polarities are heard, the higher is the RUE. In contrast to Aicher et al. [3], we determine the RUE dependent on the respective clusters (subtopics) k. The RUE increases if number of heard pro and con arguments is balanced or/and the more arguments are heard. To take a potential, data-related bias in cluster k (number of pro and con arguments unequal) into account, we introduce the characteristic function 1.

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

Throughout Section 3 s_k denotes the number of single arguments belonging to cluster k and p denotes pro/con pairs ($p_k = s_{k,\rho} \wedge s_{k,\overline{\rho}}$) of the respective cluster k (e.g. $s_1 = 3$ indicates that for the cluster 1 three single arguments exist). The index a denotes all elements in this cluster, v denotes visited and thus heard arguments, ρ denotes an argument's polarity which corresponds to the user's point of view and $\overline{\rho}$ an argument's polarity which contradicts the user's point of view. This implies, that at the beginning of the interaction, the users have to state their point of view and are furthermore able to update this information anytime during the dialogue. Without loss of generality if the user states to be indifferent, we make the conservative assumption that $s_{k,v} = s_{k,v,\rho}$[1].

Eq. (1) considers if at least one pro/con pair[2] has been heard and if so, makes it possible to take into account additionally heard single arguments. However, the latter is limited to cases where users hear additional arguments that contradict their own point of view or only a limited number of arguments that support their own point of view. Thus, we reward a balanced exploration or that of the opposite point of view more, since, as indicated before, this requires a greater effort from the user.

If there exist additional single arguments ($s_k \geqq 0$) we define these singles with $s_{k,w}$ in Equation (2). Following the same considerations as for Eq. (1) we distinguish three cases. The first describes the event that the number of visited single arguments $s_{k,v}$ is smaller than the total number of single arguments $s_{k,a}$ in cluster k. The second considers the event where the user explores more singles than pairs, which are in line with their point of view. As this does

[1]This means if users listen to single arguments and are indifferent regarding their stance, the RUE is calculated as if they were in line with the respective single arguments.
[2]A pro/con pair is defined as a pro and a con argument, regardless their relation to each other. Only their polarity with regard to the topic of the discussion is important.

not point to a critical, reflective opinion-forming, it is weighted with a downsizing factor $\gamma_k \in (0,1)$. The third case considers the case where the user explores more opposing singles than pairs, which indicates a higher RUE. The first and third case are weighted with an additional factor $\beta_{1,k}, \beta_{2,k} \in (0,1]$ which is set to 1 for all k as a starting point and can be adapted accordingly if e.g. the exploration of opposing arguments should be rewarded more.

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

where $s_{k,v}$ denotes the visited single nodes which are either in line with or oppose the user's point of view. $s_{k,a}$ describes all existing singles in a cluster (counterpart to $p_{k,a}$). The first part of the numerator of the term $s_{k,v,\overline{\rho}}$ denotes the number of visited arguments which oppose the user's point of view and belong to cluster k. The second part displays the already heard pairs $p_{k,v}$, which should not count into $s_{k,v,\overline{\rho}}$ as a respective counterpart has been heard. The denominator consists of all arguments which oppose the user's point of view $s_{k,a,\overline{\rho}}$ subtracted by $p_{k,v}$. Henceforth, Eq. (2) takes account for the fact, that the exploration of an opposing view counteracts the SFB and thus, has a higher impact, than the exploration of arguments which stress the user's point of view.

Please note that, in case only pairs and no singles exist ($s_{k,all} > 0$), Eq. (2) can be simplified to:

$$s_{k,w} = \begin{cases} 0, & \text{if } \sigma_{k,v,\rho} \geqq \sigma_{k,v,\overline{\rho}} \\ \beta_{2,k} \frac{\sigma_{k,v,\overline{p}} - p_{k,v}}{\sigma_{k,a,\overline{p}} - p_{k,v}}, & \text{if } \sigma_{k,v,\rho} < \sigma_{k,v,\overline{\rho}} \end{cases}. \tag{3}$$

It follows for the resulting RUE component r_k of the respective cluster k can therefore be determined by:

$$r_k = \alpha_k \frac{|p_{k,v}|}{|p_{k,a}|} + \mathbb{1}_{p_{k,v}} (1 - \alpha_k)\, s_{k,w}, \tag{4}$$

with $r_k \in [0,1]$. In Eq. (4) visited pairs $p_{k,v}$ are weighted with a factor α_k and all single arguments with $(1-\alpha_k)$ which leaves some room as to how much a balanced exploration is rewarded. Without loss of generality, if no pro/con pairs exist in the cluster k ($|p_{k,a}| = 0$) it follows $\alpha_k := 0$; $\frac{|p_{k,v}|}{|p_{k,a}|} := 0$.

3.1.2 Personal Relevance (PR). According to Westerwick et al. [42] which of two different information processing modes (central/peripheral) is chosen, depends also on the individual user motivation e.g. Personal Relevance. Thus, we chose the Personal Relevance (PR) as another dimension in our SFB model. The PR refers to the user individual assessment of how relevant a cluster is with regard to the topic of the discussion. Thus, each cluster is assigned a certain value, e.g. a 5-point Likert-scale rating. For instance, the user could rate the statement "This aspect is personally relevant to me in the discussion of {topic}" for each cluster: 5 = Strongly agree , 4 = Agree, 3 = Neutral, 2 = Disagree, 1 = Strongly disagree. By normalizing the obtained rating, we obtain for the personal relevance pr of the respective cluster k:

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

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

3.1.3 True and False Knowledge. Besides the previously mentioned dimensions also the ability e.g. preexisting knowledge [42] is crucial for a user to process information via the central route and thus, thoroughly scrutinizing it according to the ELM model. Building upon this argumentation, we consider the (preexisting) knowledge by distinguishing two correlated dimensions: True Knowledge (*TK*) and False Knowledge (*FK*).

The True (False) Knowledge is defined as the user's correct or respectively, incorrect knowledge on the current cluster at the current state of the interaction. The system's database does contain only validated and thus, correct information and consequently, information contradicting the former is incorrect. We aim for the user to explore as much information as possible, as this increases the chance to explore other aspects and viewpoints. Thus, the greater the user's True Knowledge, the more unlikely he/she finds themselves in an SFB concerning the respective cluster. Vice versa, if the user is misinformed on certain aspects, it increases the risk of being stuck in an SFB and being reluctant towards contradicting correct information. Therefore, the greater the false knowledge regarding a cluster, the more likely the users find themselves in an SFB. Consequently, we define the True Knowledge tk_k concerning a cluster k as the relation of the number of arguments belonging to k the user listens to during the interaction and the total number of arguments belonging to k (n_k). As we want to distinguish between the preexisting knowledge of the user and the newly gained one through the interaction, we furthermore define the initial True Knowledge $tk_{k,i}$ as the relation of the number of arguments the user states to already know ($n_{k,v,known}$) and n_k. It follows:

$$tk_k = \frac{n_{k,v}}{n_k}, \tag{6}$$

$$tk_{k,i} = \frac{n_{k,v,known}}{n_k} \tag{7}$$

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

In order to display the False Knowledge fk_k regarding arguments belonging to cluster k in the similar range $(0,1]$ as the other dimensions, we define the inverse relation:

$$fk_k = \frac{1}{1 + \theta_k n} \quad \forall n \in \mathbb{N}_0 \tag{8}$$

where n denotes the number of instances where the user stated to have contradicting information and $\theta \in (0, \infty)$ displays a weighting factor which can be chosen accordingly. Without loss of generality we choose $\theta_k = 0.5 \; \forall \; k$ as a starting point [3]. In case of $n = 0$ we define $fk_k := 1$ and thus, no False Knowledge with regard to cluster k is present. Therefore, the probability for the user to be open-minded towards presented arguments is higher.

3.2 Clusterwise SFB-Model

Using the previously defined dimensions and derived Equations (4), (5), (6) and (8) we obtain the user's SFB-vector for each

[3] As it might be useful to adjust θ_k according to the cluster sizes, we chose a ten-dimensional $\vec{\theta}$ instead of θ, as the herein discussed dataset consists of 10 clusters.

cluster k:

$$\overrightarrow{sfb_k} = (pr_k, r_k, tk_k, fk_k)^T. \tag{9}$$

In Figure 1 an exemplary sketch of this vector and the respective SFB of this 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 describe a probability for a user to be inside or outside the SFB. Depending on the definition of "breaking"

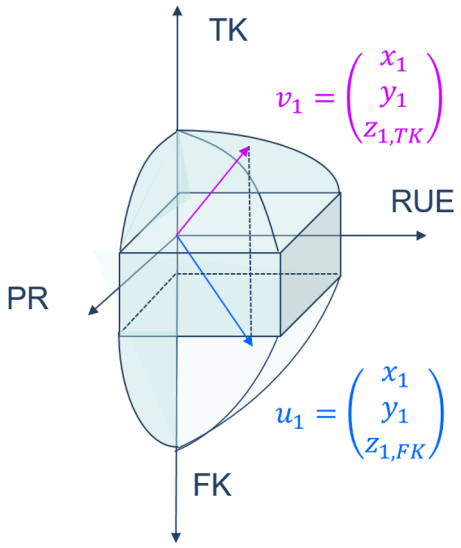

Figure 1: Schematic sketch of a clusterwise SFB-vector and SFB for a cluster k. The box indicates, the probability of a filter bubble is very dense and high near the origin and if a dimension is close to zero. For better illustration the four-dimensional SFB-vector is displayed in two split components 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.

the SFB for a cluster, using this vector various criteria could be examined. We suggest to examine in a first step the initial (before the interaction) and final (after the interaction) vector position with respect to the clusterwise SFB. For instance, a minimum relative change $\delta_{k,min}$ between the initial and final position could be defined and compared by calculating the difference in magnitudes δ_k:

$$\delta_k = |\overrightarrow{sfb_{k,f}}| - |\overrightarrow{sfb_{k,i}}|, \tag{10}$$

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

3.3 Overall SFB-Model

In order to define an overall SFB-Model for all clusters, the differences between these cluster have to be taken into account regarding the determination of the reflective user engagement. When considering hierarchical argumentation structures, e.g. argument trees, arguments at the beginning of a branch are more general than ones at deeper levels. Due to this we introduce a hierarchical weight ω_d in order to incorporate the different levels of argument depth into the overall RUE measure. Therefore, a balanced exploring of lower levels will be assigned larger weights than near the root node (see Fig. 2. As the depth of arguments within the argument tree may vary, we define a median depth $d_k = \text{med}(D)$ with D denoting all depths of the respective visited arguments belonging to cluster k[4]. Thus it follows

$$\omega_{d,k} = \frac{d_k}{\sum_{m=1}^{d_{k,max}} m}, \tag{11}$$

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

Furthermore, to avoid an over-representation of clusters with only a few arguments while clusters with many arguments will be under-represented, we define a weight $\omega_{k,n}$ which takes the different sizes of clusters into account. Thus, we relate the number of arguments n_k within the cluster k to all arguments in all clusters n_{all} such that:

$$\omega_{n,k} = \frac{n_k}{n_{all}}. \tag{12}$$

By merging the Equations (4), (11) and (12) and respective normalization, we obtain the overall RUE for all n cluster:

$$RUE = \frac{\sum_{k=1}^{n} \omega_{d,k}\omega_{n,k}r_k}{\sum_{k=1}^{n} \omega_{d,k}\omega_{n,k}}, \tag{13}$$

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

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

$$X = \frac{\sum_{k=1}^{n} x_k}{n}, \tag{14}$$

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

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

[4]Taking the average instead would lead to a great bias, especially as $d \in \mathbb{N} \quad \forall\, d \in D$

ζ_1 and ζ_2 are margins between 0 and 1, which have to be chosen according to how strict the SFB is defined. However, especially as the definition of these margins is difficult, we recommend to take also other criteria or rather combinations of criteria into account, for instance:

- Extremes: dimensions (clusterwise and overall) which are equal to 0 or 1
- Relative changes of the clusterwise SFB-vectors (initial vs. final position)
- Dimension related margins
- Great discrepancies between clusters

The above mentioned criteria could also provide useful information for the system's policy to break the SFB during the interaction. Which combination of these criteria is the most suitable shall be explored in future work and we encourage to adapt them according to the respective application setting.

3.4 Requirements: Annotation Scheme and Argument Clustering

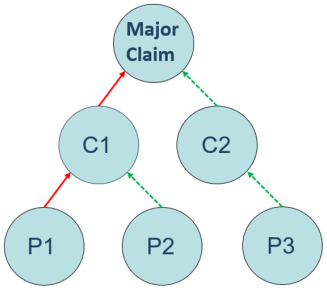

Figure 2: Visualization of argument tree structure. The topic of the discussion is the root node, which is supported by the claim C2 (green dotted arrow) and attacked by claim C1 (red solid arrow). The respective leaf nodes are the premises P1, P2 and P3.

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 [40][5]. It distinguishes three different types of components (Major Claim, Claim, Premise), which are structured in the form of bipolar argumentation trees depicted in Figure 2. The overall topic of the debate is formulated as the *Major Claim* Φ_0 representing the root node in the graph. *Claims* (C1 and C2) on the other hand are assertions which formulate a certain opinion targeting the *Major Claim* but still need to be justified by further arguments, *premises* (P1 and P2) respectively. We consider two relations between these argument components (nodes): *support* (green dotted arrows) or *attack* (red solid arrows). Each component apart from the Major Claim Φ_0 (which has no relation) has exactly one unique relation to another component. This leads to a non-cyclic tree structure, where each node or "parent" (C1 and C2) is supported or attacked by its "children". If no children exist, the node is a leaf (e.g. *P*1, *P*2 and *P*3) and marks the end of a branch.

Furthermore, our SFB model requires semantically clustered arguments, such that each argument belongs to one or more clusters of the discussed topic. There are many different approaches for clustering data. Research in argument clustering is mostly based on textual structures or linguistic features using agglomerative clustering [10, 35]. However, as an argument can address more than one aspect of a topic, it may belong to multiple overlapping clusters [13]. Thus, according to Reimers et al. [36], simple partitioning algorithms such as agglomerative clustering are unsuited for argument clustering. As machine learning techniques to identify semantic clusters are very complex, for a first implementation we will make use of manual clustering by human expert annotators and will focus on the former in future work. Due to the fact that manual clustering captures semantically fine-grained nuances it may even be better in estimating the similarity of arguments [13]. 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. Note that an argument component can both directly and indirectly address the same cluster, i.e. if it belongs to a cluster itself and it also inherits the cluster from its parent. The major claim denoting the overall topic does not belong to a cluster.

4 SFB-MODEL INTEGRATION INTO THE ADS

In the following, the relevant components of the ADS with regard to the exemplary integration of our model are introduced. After an overview of its knowledge base and argument clustering, the underlying dialogue model and interface of the ADS are described.

4.1 Knowledge Base and Argument Clustering

In this ADS a sample debate on the topic *Marriage is an outdated institution* provides a suiting argument structure and fulfills all requirements in Subsection 3.4. It serves as knowledge base for the arguments and is taken from the *Debatabase* of the idebate.org[7] website. It consists of a total of 72 argument components (1 *major claim*, 10 *claims* and 61 *premises*) and their corresponding relations and is encoded in an OWL ontology [9] for further use. In each "why pro/con" move a single argument component is presented to the user. The maximal depth of a branch d_{max,B_j} varies from 5 up to 10. To prevent the user from being overwhelmed by the amount of information, the available arguments are presented to the users incrementally on their request. The allowed moves the user is able to make are explained in Subsection 4.2. With regard to the argument clustering, the following ten clusters were identified in our sample dataset: *Alternative relationships and parenthoods, Children, Divorce, Expectations and commitment, Harmful relationships, Law, Relationship stability, Religion, Remarriage, Social Acceptance.*

[5]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.

[7]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.

Table 1: Description of the possible user moves with corresponding determiners and influenced SFB dimension. The latter is updated dynamically after each move.

Move	Description	Determiners	SFB Dim
why_{pro}	Request for a pro argument	If supporting child exists	r_k, tk_k
why_{con}	Request for a con argument	If attacking child exists	r_k, tk_k
$suggest$	Suggest another argument (random/interest)	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	$tk_{k,i}$[6]
$false$	States that current argument is incorrect	Always	fk_k
$exit$	Terminates the conversation	Always	

4.2 Dialogue Model

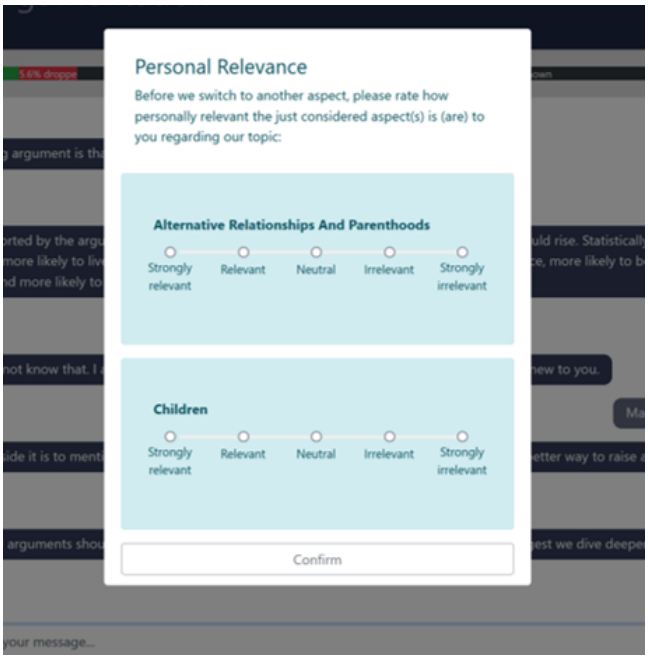

Figure 3: Exemplary Popup for the PR 5-point Likert rating of the clusters "Alternative relationships and parenthoods" and "Children" which were previously addressed before the switch to another cluster.

Besides the previously described requirements, in order to apply the SFB-Model introduced in Section 3 the dialogue model has to provide respective user moves. The interaction between the system and the user is separated in turns, consisting of a user action and corresponding natural language answer of the system. The system response is based on the original textual representation of the argument components, which is embedded in moderating utterances. Table 1 shows the required[8] 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 (root / parent node / leaf node).

As can be seen 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 form of a pop-up window which is shown in Figure 3. 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).

4.3 ADS Interface

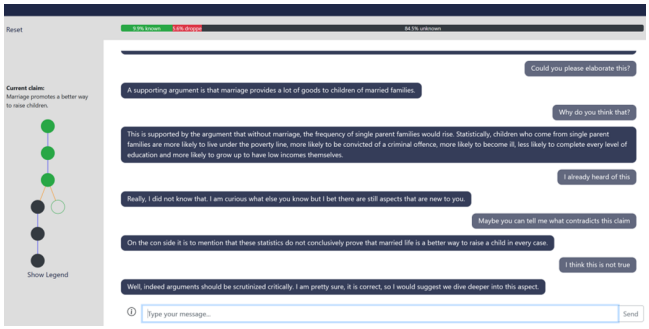

Figure 4: GUI of the ADS. Above the input line the dialogue history is shown. The graphical visualization of the current argument branch, its legend and the corresponding root node of the graph are shown left to the dialogue history. A help button left to input line provides suggestions for requests in case the user does not know how to proceed.

The graphical user interface (GUI) of the ADS is illustrated in Figure 4. The users type their request (action) in the input line below the dialogue history and press "Send". This input is interpreted by an NLU framework [1] which processes the typed user utterance using an intent classifier based on a BERT Transformer Encoder [15] and a bidirectional LSTM classifier. After a user move is recognized, the dialogue management reacts accordingly and the corresponding system's response is displayed on the screen. To make sure that omitted arguments were skipped deliberately,

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

a graphical visualization left to the dialogue history in Figure 4 shows the users the length and structure of the currently visited argument branch. It shows the current root of the argument branch, the argument branch itself and the user's current position (green bordered node) are displayed. Already visited arguments are shown in green and unheard ones in blue.

In order to test the validity and consistency of our SFB Model in a proof-of-principle scenario, it was integrated into this cooperative ADS which served as the study setting described in the following.

5 USER STUDY

We conducted a user study with 102 participants divided into two groups, an experimental group with interest-driven choice of arguments (hereinafter referred to as "interest") and control group with random choice (hereinafter referred to as "random"), to show the validity of the proposed SFB-Model. For the interest group the ADS chose arguments which suited the user's preferences (shown by preference or rejection moves of the user) and interest (modelled by an interest model [2]) best, whereas the random group received randomly chosen arguments. Thus, the interest group was always presented arguments of the requested polarity and cluster, whereas the arguments for the random group were picked randomly from the list of still available (unheard) arguments. The study aimed for analyzing the following research questions:

(1) Is the presented model suitable to describe a user's SFB?
(2) Do the SFB dimensions change dynamically and if so do these changes match the expectations based on the course of the dialogue?

More precisely, we defined the following hypotheses to be examined in the study:

H1 Participants in the interest group showed a higher probability to be caught in an SFB.
H2 The changes in the SFB dimensions are consistent with the expected behaviour based on the course of the dialogue.

The study was conducted online via the crowdsourcing platform "Crowdee" (https://www.crowdee.com/) with participants from the UK, US, and Australia (English native speakers to avoid language barrier effects). The duration of the interaction was estimated to be about 20 minutes which was rewarded with 5.80€ (17.40€/hour). The study setup used the chat-based output modality. After an introduction to the system (short text and description of how to interact with the system) the users had to pass two control questions to check whether they understood how to interact with the system. The users who passed this test, were advised to explore enough arguments to build a well-founded opinion on the topic *Marriage is an outdated institution*. The participants were not told anything about the underlying SFB-Model or Interest-Model but only to request at least ten arguments. In addition, they were asked to rate their opinion and interest on the topic on a 5-point Likert-scale, which was normalized in $[0, 1]$. During and after the study, we collected the following data[9]: Self-assessment (questionnaire) and calculated values for RUE, PR, TK, FK for each cluster k, opinion and interest on the topic of the discussion, the set of heard arguments, the dialogue history.

We conducted this user study with 102 participants aged 19-36 (average age: 36.3 (SD 9.1)). The interest group comprised 53 users (35 females, 17 males, 1 other/do not want to tell) and the random group 49 (33 females, 14 males, 2 other/do not want to tell). All participants were non-experts with no topic-specific background. Both groups did not differ in their experience with CUIs (Interest: 2.40 (SD 1.12), Random: 2.37 (SD 1.21) on the 5-point Likert-scale from 1 - "No experience" to 5 "Very much experience").

6 RESULTS

On average, participants spent 37.77 min in the interaction with the system (interest: 36.51 min, random: 39.15 min). Most of the participants (interest: 55%; random: 68%) heard 20-30 out of 72 available arguments. A total of 7 participants (6.8%, interest: 5, random: 2) quit the interaction after the minimum number of ten presented arguments was reached. These findings already indicate that the saturation effect is stronger in the interest group as the users are likelier to consume only the arguments which suit their interest best and then quit the system [10].

Table 2: Means and standard deviations of all SFB dimensions over all cluster for both groups. Statically (very) significant differences are indicated in bold p values $p < 0.05$ ($p < 0.01$). Furthermore the respective effect size is displayed.

	Interest		Random			
Asp.	*M*	*SD*	*M*	*SD*	*p* value	effect size *r*
RUE	0.205	0.254	**0.321**	0.313	**0.003**	0.314
PR	0.763	0.223	**0.764**	0.236	0.306	0.035
TK	0.337	0.221	**0.512**	0.356	**<0.001**	0.452
FK	0.863	0.103	**0.924**	0.122	**0.034**	0.113

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 only show the weighted overall mean for each SFB dimension averaged over all clusters. As the differences between the cluster are very big and vary in size, the respective standard deviations are large but comparable between groups. Furthermore, the general observations we describe here also apply to the individual mean values of the SFB dimensions for each cluster. Strikingly, RUE, TK, and FK (inverted!) are significantly[11] larger for the random group[12]. Specifically, the largest and most significant difference was observed in TK where the Pearson's correlation coefficient (effect size) was found to be medium ($0.3<r=0.452<0.5$) with a p-value < 0.001. Also the significant difference ($p=0.003$) in the RUE dimension is of medium effect size $r=0.314$. Albeit with a small

[9] All regulations regarding data protection and anonymity of users were strictly adhered to at all times, and the participants were able to quit the study at any time. An ethical review by an IRB was nor required due to internal guidelines due to the solely cooperative, non-persuasive design of the user study.

[10] The longer the ongoing interaction, the more likely it is, that the system will present less suiting arguments, as they are no better suiting arguments available anymore.

[11] To determine whether the difference between the two group means is significant, we used the non-parametric Mann-Whitney U test for two independent samples with no specific distribution as the values are not normally distributed according to the Shapiro Wilk test.

[12] Please note, that higher numbers are related of a higher probability to be outside the SFB

effect size (r=0.113), the difference in FK was found to be significant (p=0.034<0.05).

While no significant difference was observed in the overall PR (all clusters) of PR, the examination of single cluster revealed that in 30% of the clusters significant differences between the two groups were noticeable ($0.037 \leqq p \leqq 0.048$, $r \leqq 0.1$).

Regarding the "pre-interest" (before the interaction) of the participants the difference between the two groups is insignificant (interest: 3.34 (SD 1.04), random: 3.53 (SD 1.17); p=0.344). Likewise, with respect the difference in their "pre-opinions" between the two groups is insignificant (interest: 2.87 (SD 1.08); random: 2.94 (SD 1.21); p=0.837). During the interaction about 20.2% (11 of 49) participants changed their opinion (from pro to con or vice versa) during the interaction in the random group and 9.4% (5 of 53) in the 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 Section 5).

7.0.1 Validity (H1): The significant differences in the overall dimensions RUE, FK, TK between both groups can be explained by the large difference in the amount of heard arguments and their corresponding polarity. Whereas the interest group was only presented with the arguments of requested polarity and the estimated most interesting cluster, the random group was presented randomly chosen arguments that did not correspond to the interest/preference of the user. Our analysis showed that, 38% more opposing arguments were displayed in the random group than in the interest group. Strikingly, when users in the interest group visited opposing arguments, they stated them to be false significantly more often (p=0.007, r=0.278) than in the random group. These observations reinforce the hypothesis that users are prone to stay within their SFBs during the exploration of controversial topics in the context of argumentative interactions with CUIs.

The overall means for PR hardly differ for both groups which indicates that the assessment of the cluster relevance seems to be independent from the change of interest of the users. Still in single clusters the significant difference correlated with lower FK values and rejection of presented arguments for the interest group. Thus, it seems users tend to rate clusters containing arguments they disagree with or misconstrue as personally less relevant.

These findings imply that our SFB Model can detect differences in the exploration behavior of users in the respective model dimensions. Meeting our expectation, our SFB Model showed a higher probability for participants in the random group not to be caught in an SFB and seem to be suitable model for SFBs of users.

7.0.2 Consistency (H2): During the interactions it was noticeable, that in 85% of the cases where users changed their opinion on the topic the respective RUE was very high, indicating a balanced or rather opposing exploration behavior. On the other hand, the participants with the lowest RUE values (below 0.15) correlated with users stated the incorrectness of an argument (73% of all $false$ moves). This implies that the RUE seems to be a strong indicator for the probability to be stuck in an SFB and matches the definition of the RUE. Due to the construction of our Model, we perceive a strong alignment between high RUE and TK, especially if the users changed their opinion. Moreover, it is noticeable that the personal relevance of clusters is more likely to increase the more arguments of the respective clusters are heard (this is especially significant for small clusters like "Law"). Thus, we can deduce that our expectations match the dynamic changes noticeable in the SFB dimensions during the course of the dialogue. This also underlines the importance of promoting a diverse range of viewpoints and encouraging the exploration of multiple perspectives in order to prevent the formation of SFB and to foster a more critical scrutinizing of complex issues.

Our results imply that we can determine if the user has a high change to be stuck within an SFB and use this information to determine which potential argument with respect to the SFB dimensions and clusters should be suggested to the user to increase the chance of breaking it. As CUIs offer a more engaging and personalized user experience compared to traditional graphical user interfaces, they are of particular interest when it comes to introducing such an argument, which has not been initially requested by the user. For instance, apart from a transparent explanation, a graphical/visual as well as different multi-modal cues.

8 LIMITATIONS

However, the previously described study is subject to two limitations that could be addressed in future research. First, the user study is only tested on one topic ("Marriage is an outdated institution"). We chose this topic as its dataset meets our requirements of being large enough, balanced (regarding argument stance pro/-con), overlapping clusters, high quality and argument depth. Even though it seems suitable for a proof-of-principle study, the generalizability of our findings needs to be shown with respect to other topics. Second, we focus on quantitative data analysis only. While quantitative data can provide useful insights, it may not capture the full range of experiences and perspectives of the participants. Therefore, in future studies, it may be beneficial to supplement the study with qualitative data in the form of participant interviews. Especially when incorporating explicit argument suggestions to break the SFB this will provide an insight how this influences the user's satisfaction and perception of the CUI.

9 CONCLUSION AND FUTURE WORK

Complex interactions with CUIs are highly likely to become even more pervasive in the future, and thus, also they will also play an important role in reducing the probability of Self-imposed Filter Bubbles in user interactions. In this work, we introduced a novel model for the SFBs of users, consisting of four dimensions: Reflective User Engagement, Personal Relevance, True Knowledge and False Knowledge (but not limited thereto). To the best of our knowledge, this model represents the first approach to estimate the probability that users find themselves within a Self-imposed Filter Bubble. After describing the choice of the main four SFB dimensions, we introduced the clusterwise (subtopic-related) and the overall SFB model, as well as approaches to detect whether the SFB of a user is broken during a cooperative argumentative dialogue. Moreover, we discussed an exemplary integration of our model

into an ADS and validated it in an online user study. The described study gave an insight into the changes of each dimension during the interaction. The results showed that our model behaves as expected and is suitable to capture the SFB of a user. In particular, it estimates a higher probability that the user is caught in an SFB if the ADS only suggests arguments, which suit the user's interest best. On the other hand, the group which was provided with randomly chosen arguments showed a lower probability of the user being stuck in an SFB which is consistent with our expectation. In conclusion, we showed that the herein presented model provides a suitable way to determine whether a user is caught in an SFB and to describe the dynamics of the SFB within an ongoing argumentative dialogue.

In future work, we want to address already mentioned open questions, for instance how to choose corresponding weights for different clusters and define margins and areas for the SFB probability. Furthermore, it shall be examined how the SFB model can be merged with other models (user interest, preference) to maintain the user's motivation to interact with the system for as long as possible. Therefore, Reinforcement Learning approaches shall be explored which enable us to adapt to the individual user and engage them to recognize and overcome their SFB. In conclusion, the herein presented model takes us a step closer to our aim to provide a cooperative ADS that helps users to build a well-founded opinion and fosters critical, reflective thinking and open-mindedness.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Waheed Ahmed Abro, Annalena Aicher, Niklas Rach, Stefan Ultes, Wolfgang Minker, and Guilin Qi. 2022. Natural language understanding for argumentative dialogue systems in the opinion building domain. *Knowledge-Based Systems* 242 (2022), 108318. https://doi.org/10.1016/j.knosys.2022.108318

[2] Annalena Aicher, Nadine Gerstenlauer, Wolfgang Minker, and Stefan Ultes. 2022. User Interest Modelling in Argumentative Dialogue Systems. In *Proceedings of the 13th Language Resources and Evaluation Conference*. Marseille, France, 127–136.

[3] Annalena Aicher, Wolfgang Minker, and Stefan Ultes. 2021. Determination of Reflective User Engagement in Argumentative Dialogue Systems. http://ceur-ws.org/Vol-2937/paper1.pdf

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

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

[6] Hassan Alsharif and John Symons. 2021. Open-mindedness as a Corrective Virtue. *Philosophy* 96, 1 (2021), 73–97. https://doi.org/10.1017/S0031819120000352

[7] Bharat N Anand. 2021. The US media's problems are much bigger than fake news and filter bubbles. *Domestic Extremism* (2021), 138.

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

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

[10] Filip Boltužić and Jan Šnajder. 2015. Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, 110–115. https://doi.org/10.3115/v1/W15-0514

[11] 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*, Jiřina Vejnarová and Nic Wilson (Eds.). Cham, 59–73.

[12] Lisa A. Chalaguine and A. Hunter. 2020. A Persuasive Chatbot Using a Crowd-Sourced Argument Graph and Concerns. In *COMMA*. https://doi.org/10.3233/FAIA200487

[13] 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 (2020), 115–121.

[14] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Sci Rep* 7, 40391 (2017), 1–9. https://doi.org/10.1038/srep40391

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

[16] 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* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 12–22. https://doi.org/10.1145/3460231.3474261

[17] 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 (2022), 100226. https://doi.org/10.1016/j.chbr.2022.100226

[18] Bettina Fazzinga, Andrea Galassi, and Paolo Torroni. 2021. An Argumentative Dialogue System for COVID-19 Vaccine Information. In *Logic and Argumentation*. Cham, 477–485.

[19] Hans Gelter. 2003. Why is reflective thinking uncommon. *Reflective Practice* 4, 3 (2003), 337–344. https://doi.org/10.1080/1462394032000112237

[20] 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 (12 2022), 1–53. https://doi.org/10.3233/AAC-210005

[21] 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 (2012), 438–447.

[22] Martin Jones and Robert Sugden. 2001. Positive confirmation bias in the acquisition of information. *Theory and Decision* 50, 1 (2001), 59–99.

[23] Andreas Kappes, Ann H. Harvey, Terry Lohrenz, P. Read Montague, and Tali Sharot. 2020. Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience* 23, 1 (2020), 130–137.

[24] 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* (2018), 121–130. https://doi.org/10.18653/v1/W18-5215

[25] Terry Lee. 2019. The global rise of "fake news" and the threat to democratic elections in the USA. *Public Administration and Policy* (2019).

[26] Robyn Macpherson and Keith E. Stanovich. 2007. Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences* 17, 2 (2007), 115–127. https://doi.org/10.1016/j.lindif.2007.05.003

[27] 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 (2020), 102939.

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

[29] Lien Michiels, Jens Leysen, Annelien Smets, and Bart Goethals. 2022. What Are Filter Bubbles Really? A Review of the Conceptual and Empirical Work. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

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

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

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

[33] 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*. Routledge, 141–180.

[34] Walter Quattrociocchi, Antonio Scala, and Cass R Sunstein. 2016. Echo chambers on Facebook. *Available at SSRN 2795110* (2016).

[35] 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*. 45–52.

[36] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and Clustering of Arguments with Contextualized Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 567–578. https://doi.org/10.18653/v1/P19-1054

[37] Ariel Rosenfeld and Sarit Kraus. 2016. Strategical Argumentative Agent for Human Persuasion. In *ECAI'16* (The Hague, The Netherlands). 320–328. https://doi.org/10.3233/978-1-61499-672-9-320

[38] 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 (2012), 2280–2290.

[39] 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 (2021), 379–384. https://doi.org/10.1038/s41586-021-03215-w

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

[41] 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 (2016), 167–179.

[42] Axel Westerwick, Benjamin K. Johnson, and Silvia Knobloch-Westerwick. 2017. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs* 84, 3 (2017), 343–364. https://doi.org/10.1080/03637751.2016.1272761 arXiv:https://doi.org/10.1080/03637751.2016.1272761