Towards Healthy Engagement with Online Debates

An Investigation of Debate Summaries and Personalized Persuasive Suggestions

Alisa Rieger
a.rieger@tudelft.nl
Delft University of Technology
Delft, Netherlands

Qurat-ul-ain Shaheen
qurat@iiia.csic.es
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain

Carles Sierra
sierra@iiia.csic.es
Artificial Intelligence Research Institute (IIIA-CSIC)
Bellaterra, Spain

Mariët Theune
m.theune@utwente.nl
University of Twente
Enschede, Netherlands

Nava Tintarev
n.tintarev@maastrichtuniversity.nl
Maastricht University
Maastricht, Netherlands

ABSTRACT

Online debates allow for large-scale participation by users with different opinions, values, and backgrounds. While this is beneficial for democratic discourse, such debates often tend to be cognitively demanding due to the high quantity and low quality of non-expert contributions. High cognitive demand, in turn, can make users vulnerable to cognitive biases such as confirmation bias, hindering well-informed attitude forming. To facilitate interaction with online debates, counter confirmation bias, and nudge users towards engagement with online debate, we propose (1) summaries of the arguments made in the debate and (2) personalized persuasive suggestions to motivate users to engage with the debate summaries. We tested the effect of four different versions of the debate display (without summary, with summary and neutral suggestion, with summary and personalized persuasive suggestion, with summary and random persuasive suggestion) on participants' attitude-opposing argument recall with a preregistered user study (N = 212). The user study results show no evidence for an effect of either the summary or the personalized persuasive suggestions on participants' attitude-opposing argument recall. Further, we did not observe confirmation bias in participants' argument recall, regardless of the debate display. We discuss these observations in light of additionally collected exploratory data, which provides some pointers towards possible causes for the lack of significant findings. Motivated by these considerations, we propose two new hypotheses and ideas for improving relevant properties of the study design for follow-up studies.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; • **Information systems** → **Personalization**.

UMAP '22 Adjunct, July 4–7, 2022, Barcelona, Spain

ACM ISBN 978-1-4503-9232-7/22/07.
https://doi.org/10.1145/3511047.3537692

ACM Reference Format:
Alisa Rieger, Qurat-ul-ain Shaheen, Carles Sierra, Mariët Theune, and Nava Tintarev. 2022. Towards Healthy Engagement with Online Debates: An Investigation of Debate Summaries and Personalized Persuasive Suggestions. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3511047.3537692

1 INTRODUCTION

Online debates, whether on designated debate platforms, in comments under news websites, or in social media comments, offer great potential for democratic discourse and debate. They provide low-threshold access to participate and the potential of far reach, breaking bounds of physical limitations. However, low-threshold accessibility can also harm online debates. User-generated contributions that are low in quality and high in quantity impede fellow users from effortlessly grasping the scope and making sense of the debate. They also impede users from making meaningful contributions to the debate and forming well-informed attitudes. Moreover, navigating and processing such contributions is cognitively demanding, leaving users vulnerable to cognitive biases [10, 11, 24, 33]. One type of bias that was found to interfere with healthy engagement[1] with online debates is the *confirmation bias* [18, 29, 30]. *Confirmation bias* describes the human tendency to favor information that confirms prior attitudes and beliefs when searching for, interpreting, and recalling information [22]. This behavior has not only negative consequences on the quality of online debates but also on individuals' attitude forming and decision making, likely being a driver of ideological polarization and extremism [15, 19].

For online debates to live up to their full potential and prevent cognitive biases such as the confirmation bias, approaches that facilitate understanding and taking in the arguments made in the debate might offer a suitable solution. This could be achieved with *nudging* by means of argument summaries [1, 21]. Nudging refers to the process of subtly modifying the choice architecture that alters people's behavior to support optimal decision making [4, 27, 32]. To further increase the effect of nudging approaches, Kaptein et al. [16] propose adaptive persuasion by personalizing persuasive messages

[1]We define *healthy engagement* as users engaging impartially with arguments that express different perspectives and forming/changing their attitude according to the quality of those arguments

based on the users' susceptibility to influence strategies, defining their *persuasion profile* [17].

These considerations and observations motivate our research into **(1) summaries** of arguments made in contributions to online debates with **(2) personalized persuasive suggestions** that aim at persuading users to interact with those summaries to support healthy engagement with online debates. This project specifically aims at testing whether summaries and personalized persuasive suggestions support users to effortlessly, accurately, and in an unbiased manner, understand and take in the arguments expressed by contributions to a debate. For this, we conducted a preregistered[2] between-subjects user study in which participants were presented with an online debate page with 18 contributions that express three supporting and three opposing arguments on a debated topic in one of four *debate display* conditions: (1) without summary, (2) with summary and neutral suggestion to engage, (3) with summary and personalized persuasive suggestion to engage, (4) with summary and random persuasive suggestion to engage. We measured the recall of attitude-opposing arguments by asking all participants whether a given argument was or was not made in the debate. We did not find evidence for an effect of either summaries or personalized suggestions on attitude-opposing argument recall. In addition, we did not observe confirmation bias in participants' argument recall. Additionally collected exploratory data points us towards assuming that properties of the study design might explain the lack of significant findings. We discuss study design modification ideas for follow-up studies, new hypotheses, and ethical considerations regarding the proposed approach.

2 RELATED WORK

2.1 Quality and dynamics of online debates

The aspects of low-threshold accessibility and resulting non-expert contributions to online debates may affect the quality of the debate negatively, e.g., through a high number of contributions that are impulsive, hostile, or bring no value to the debate, instead of being rational, respecting, and constructive to the debate [2, 6, 8, 9]. The high cognitive demand of navigating, understanding, and taking in large amounts of low-quality contributions in a limited amount of time could lead to information overload, leaving users vulnerable to cognitive biases [10, 15, 24, 25, 33]. Cognitive biases are known to negatively impact individuals' attitude forming and decision making, likely driving ideological polarization and extremism [15, 19].

Two types of bias that were found to hinder healthy engagement with online debates are the *confirmation bias* and the closely related *disconfirmation bias* [18, 29–31]. Whereas the *confirmation bias* describes the human tendency to favor information that confirms prior attitudes and beliefs when searching for, interpreting, and recalling information [22], *disconfirmation bias* describes the tendency to evaluate evidence biased by prior attitudes and discount evidence that counters prior attitudes [30]. Karlsen et al. [18] investigated both biases with respect to participants' attitude change after engaging with an online debate with an extensive survey and a user study. They found evidence for *confirmation bias* and *disconfirmation bias*, with increased attitude reinforcement effects for users with a strong prior attitude. The authors concluded that the dynamics of online debates could thus be described as *trench warfare*, implying that users do interact with users who hold different values and attitudes (as opposed to the dynamic of *echo chambers* in which attitudes are thought to be reinforced due to a lack of interaction with opposing arguments), but that their prior attitudes are still reinforced by these interactions. Karlsen et al. [18] mention, however, that whether *trench warfare* poses a problem depends on the normative perspective. If for example the objective is that different viewpoints are exchanged and engaged with, *echo chambers* would pose a greater problem state than exposure to attitude-opposing arguments, even though causing short-term attitude reinforcement, might still cause long-term learning and attitude change. Taking this consideration into account, we want to investigate confirmation bias during argument recall with this study. Accurate argument recall demonstrates that users sufficiently engaged with the contributions to the debate. Thus, we assume that the foundation for long-term learning and attitude change would be achieved for users who accurately recall the (attitude-opposing) arguments of a debate.

2.2 Cognitive bias mitigation

To mitigate users' cognitive biases, such as *confirmation* and *disconfirmation bias* during interactions with online debates, Lorenz-Spreen et al. [21] suggest redesigning the digital environment with interventions aiming at behavioral change. For this purpose, they propose either *nudging* [32], an alteration of the choice architecture to support unbiased behavior and decision making, or *boosting* [14], an attempt to teach and empower users to become resistant to cognitive biases. The relatively new approach of *boosting* has been found to effectively increase users' resilience to various pitfalls of web interactions (i.e., fake news, microtargeting, confirmation bias during search) [20, 23, 26].

A suitable tool to reduce cognitive load are nudges such as argument summaries that *facilitate* understanding [4]. While recent research found that summaries cause improvements in users *sense making* (accuracy of the internal representation of the overall debate) and perceived *quality of debate* [1], we are not aware of any research that has tested the effect of debate summaries on argument recall. The second boosting element that we want to investigate is inspired by the work of Kaptein et al. [16], who found that tailoring messages according to a person's persuasion profile increases the effect of the persuasive message on the person's behavior. We thus want to investigate whether we can observe a similar effect in the context of cognitive bias mitigation during interaction with online debates. Specifically, we want to investigate the effect of personalized persuasive suggestions to engage with attitude-opposing and attitude-confirming arguments in a debate summary on argument recall.

Following the previous findings and consideration, we expect that summaries and personalized persuasive suggestions effectively mitigate confirmation bias when engaging with an online debate and formulated the following hypotheses:

(H1): Participants who see a summary based on arguments expressed by the contributions to the debate will perform better for

[2]Preregistering meant publicly determining our hypotheses, experimental setup, and analysis plan before any data collection. The (time-stamped) preregistration document can be found in our repository: https://osf.io/uv48w/.

attitude-opposing questions in the argument recall task than participants who do not see a summary.

(H2): Participants who see a personalized summary suggestion will perform better for attitude-opposing questions in the argument recall task than participants who do not see a summary and who do see a summary without a personalized summary suggestion.

3 METHODS

To test **H1** and **H2**, we conducted a preregistered user-study with a between-subjects design. We compared participants' attitude-opposing (AO) argument recall after engagement with an online debate between four groups in which participants were presented with different versions of the debate interface. Thus, we manipulated the independent variable **debate display** (see Figure 1: *without summary, with summary and neutral suggestion, with summary and personalized persuasive suggestion, with summary and random persuasive suggestion*) and measured the dependent variable **AO argument recall**. This dependent variable describes the proportion of attitude-opposing arguments that participants correctly recalled in the argument recall test (for details see Section 3.2).

3.1 Sample

We anticipated to observe a moderate effect (Cohen's $f = 0.25$) for the factor of **debate display** on participants' **AO argument recall**. With a Bonferroni-corrected significance threshold of $\alpha = \frac{0.05}{2} = 0.025$, and a power of $(1-\beta = 0.8)$, an a-priori power analysis resulted in the required sample size of 211 participants. We recruited participants via the online participant pool *Prolific*[3]. Participants who completed all three parts of the study received a reward of £2.20 for their participation. We ran the study in multiple batches which were launched at different times of the day to increase the diversity of the sample by allowing for participation from different time zones. Participants were required to be fluent English speakers, at least 18 years old, and could only participate once. All participants provided their informed consent to participate in the study and data handling.

We initially recruited 284 participants of which 70 only participated in the first part of the study and thus were excluded from data analysis since they either did not hold a strong attitude on any of the topics, or only held strong attitudes for topics and stances for which the quota of 36 participants was already reached. Those 70 participants received a partial payment of £0.25 for their effort. Out of the 214 remaining participants, we excluded the data of two participants who spent less than 30 seconds on the debate page (see Figure 1), consistent with our preregistered exclusion criterion. We did not have to exclude any data for failed attention checks, since none of the participants failed two or more of the five attention checks. None of the participants decided to withdraw their participation when given the option after the debrief (see procedure in Section 3.3). Thus, our final dataset contains data of **212 participants** of which 50% reported to be male, 47.6% female, 1.4% non-binary/other, and 0.9% preferred not to report their gender. Concerning the age of the participants, 51.8% reported to be between 18 and 25, 30.1% between 26 and 35, 12.2% between 36 and 45, 4.2% between 46 and 55, 1.4% between 56 and 65 years old.

[3]Prolific: https://www.prolific.co/

3.2 Material

Topics, Arguments, and Contributions. The contributions and arguments we displayed on the debate page were sampled from the *IBM ArgKP dataset* [3] collected as part of the *Debater project*.[4] The *IBM ArgKP dataset* contains 24093 statements for 28 debated topics. The statements were collected actively from crowdworkers for the *IBM ArgQ Rank 30kArgs* dataset [12], are limited in length to 35 to 210 characters and had to fulfill quality control measures. This dataset further contains 4-11 keypoints that summarize the arguments made by the crowdworkers for each of the 28 topics. Statements that argue in line with a given keypoint are labeled as such. Further, they are labeled as either supporting or opposing the topic statement. From this dataset, we selected three topics by applying inclusion criteria regarding the number of keypoints and the balance of attitudes of 100 crowdworkers in a pretest (see preregistration in our repository at link in footnote 2). These criteria resulted in the following three topics: (1) We should legalize sex selection; (2) We should abandon the use of school uniforms; (3) We should abolish capital punishment. We randomly selected three supporting and three opposing keypoints for each topic that were displayed as the summary in the **debate display** conditions with summary (see "Debate Summary"). For each of these six keypoints, we randomly sampled three statements labeled as making an argument in line with a given keypoint. To facilitate the evaluation of the argument recall test, we did not sample any statements that were labeled as being in line with more than one keypoint. The resulting 18 statements per topic were displayed as contributions (see Figure 1).

Susceptibility for Persuasion Scale (STPS). The *STPS* scale consists of a 26-item questionnaire[5] that measures a person's susceptibility to the six distinct social influence strategies [16]. The strategy participants score highest in is considered to be their *persuasion profile* [17]. In line with [16], participants in the *personalized persuasion* condition were presented with a randomly selected summary suggestion out of the three suggestions we formulated (see *summary suggestions*) for their *persuasion profile*.

Debate Summary. Figure 1 shows an example of the debate page with summary. In the display condition *with summary*, participants were presented with a summary suggestion (see next paragraph) and four keypoints (two attitude-supporting, two attitude-opposing). Participants had the option to unfold two additional keypoints (one supporting, one opposing) when clicking on a *show more* button. This feature allowed us to explore the effect of the persuasive suggestions on the engagement with the summary with a direct behavioral measure.

Summary Suggestions. We formulated different versions of the summary suggestions that we displayed on top of the summary (see Figure 1) following Cialdini's six principles for persuasion [5] and the process of designing persuasive messages applied by [16]. For each of the six persuasion principles we formulated three versions. We tested these versions with a mapping task, in which we asked four colleagues to map each of the statements to one or more

[4]IBM Debater project: https://research.ibm.com/haifa/dept/vst/debating_data.shtml
[5]The questionnaire with all 26 items can be found in the preregistration in our repository at the link in footnote 2.

persuasion principles. The evaluation criteria and the final 19 versions of the summary suggestion can be found in the preregistration in our repository at link in footnote 2.

Argument recall Test. With the argument recall test, we aimed at measuring what contributions participants engaged with on a deep enough level to remember the argument this contribution expressed after the interaction. During the argument recall test we displayed slightly modified versions (to avoid direct recognition from the summary) of all ten sampled keypoints, including the four keypoints that were not expressed by the contributions, one after another in random order. For each keypoint we asked the following question: *In the contributions to the debate you just saw, did someone make an argument in line with the following argument: MODIFIED KEYPOINT.* Participants could answer this question by selecting either *yes*, *no*, or *I don't remember*. From the participants' responses, we calculated the dependent variable **AO argument recall** (percentage of all correct responses for attitude-opposing keypoints, three *yes*, two *no*) and the overall rate of accurate responses (exploratory, six *yes*, four *no*).

3.3 Procedure

We collected the data in three steps, approved by the ethics committee of our institution: (1) introduction and a pre-interaction questionnaire, (2) presentation of and interaction with the debate interface, and (3) post-interaction questionnaire. For the data collection we used the survey tool *Qualtrics*.[6] We ensured good data quality by including five attention checks, three in the pre-interaction and two in the post-interaction questionnaire, in which we told participants which response option to select. To ensure a balanced distribution of supporting and opposing attitudes for the three topics in our user study, we included quotas for the numbers of participants for the different topics and stances (supporting, opposing). The quota for each topic and stance was set at 36 participants (rounded up from $\frac{1}{6}$ of 211, the required number of participants). If participants only reported to have strong attitudes on topics with the stance for which the quota was already fulfilled, or did not report to have a strong attitude on any of the topics, they were not allowed to participate further and received a partial reward of £0.25, proportional to the time they invested.

Pre-interaction questionnaire: Participants were asked for their demographics and to state their attitude on the three selected topics on a seven-point Likert scale ranging from "strongly disagree" to "strongly agree". After assigning them to a topic, we measured participants *persuasion profile* with the *STPS* scale.

Interaction with the debate interface: We displayed the following instruction for the task: *Imagine you have a discussion on TOPIC with a colleague. You interrupt the discussion because your colleague has to go to a meeting, but you agree to continue the discussion later that week. To find additional arguments, you conduct an online search which leads you to the debate platform you will see after clicking "start".*

After clicking on *start* they proceeded to the debate page with 18 statements (three statements for each of the six randomly selected keypoints), and, depending on the display condition, a summary and summary suggestion. Participants were free to choose how much time they wanted to spend on this debate page and had the option to contribute to the debate. We included this option to better assimilate real online debate pages and to explore whether debate summaries might affect participants' motivation to contribute to the debate.

Post-interaction questionnaire: After the participants completed the exploration of the debate page, we asked them to complete the *argument recall test*. Then, we debriefed participants and explained the purpose of the *argument recall test*, since they only consented to "measuring their engagement" in the informed consent to avoid altering engagement with the debate page and gave the option to withdraw their participation. Lastly, participants could give feedback on the debate interface and the experimental task.

[6]Qualtrics: https://www.qualtrics.com/uk/

4 RESULTS

In the following, we describe the results of this user study, starting with a description of the dataset. Next, we present the results of testing **H1** and **H2**. Lastly, we present descriptive analyses of the exploratory data we collected in addition to the data required for testing the two hypotheses.

4.1 Description of the dataset

The participants' feedback for the task was overwhelmingly positive. Participants' high motivation for and enjoyment of the task is further supported by the mean time of 3.43 minutes they spent on the debate page. Further, 81% of all participants made a contribution to the debate. The high quality of submissions is further indicated by none participant failing more than one attention check. The number of participants per topic and stance was, in consistence with the quota we set, approximately equally distributed with 34 participants each in the group of participants who supported *We should legalize sex selection* and who opposed *We should abolish capital punishment* and 36 participants in each of the four remaining groups of topic-stance combinations. The distribution of participants in the four *debate display* conditions was likewise approximately equally distributed with (1) 25%, (2) 26%, (3) 26%, and (4) 23%, who (1) saw the debate without summary, (2) with summary and neutral suggestion, (3) with summary and personalized persuasive suggestion, and (4) with summary and random persuasive suggestion, respectively. Results of the *STPS* showed that the distribution of persuasion profiles (highest scoring persuasion category) across participants was far from being equal in our sample, with most participants scoring highest in either reciprocity (32%), commitment (27%), or liking (27%).

4.2 Hypotheses testing

Consistent with the preregistration of this user study, we tested the hypotheses with a one-way ANOVA. The result of this ANOVA shows no difference in participants' *AO argument recall* between the different *debate display conditions*. Thus, we did neither find evidence for **(H1)** an effect of summaries, nor for **(H2)** an effect of personalized persuasive suggestions on *AO argument recall* ($F(3, 208) = 0.65, p = .59$) (see Figure 2).

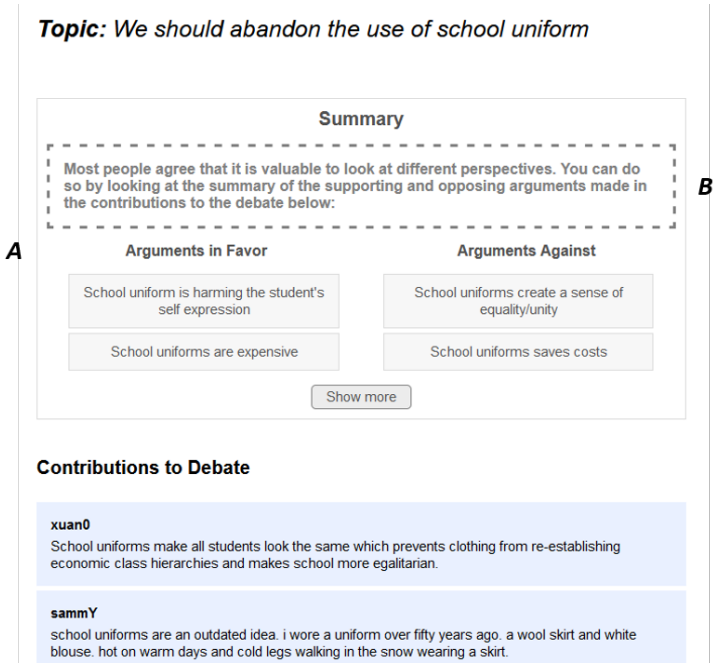

Figure 1: Debate page: ***(Example for debate display condition (2), with summary (A) and personalized persuasive suggestion (B) based on consensus strategy, and two out of 18 contributions.)***

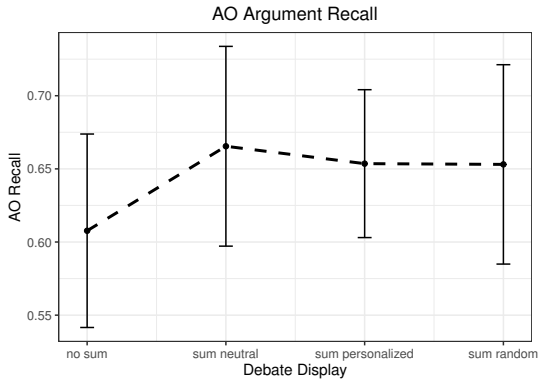

Figure 2: AO argument recall per debate display condition: mean proportion of correctly categorized arguments in the argument recall test for the debate display conditions without summary (no sum), with summary and neutral suggestion (sum neutral), with summary and personalized persuasive suggestion (sum personalized), with summary and random persuasive suggestion (sum random) with 95% confidence intervals.

4.3 Exploratory observations

The following results are exploratory, implying that we did not preregister any hypotheses or data analysis and merely look at descriptive statistics. These observations should thus not be mistaken for stand-alone findings but considered as exploratory insights that help us to better understand our observations concerning **H1** and **H2**, and to formulate new hypotheses and propose research design ideas for follow-up studies.

Confirmation bias during argument recall In the collected data we did not observe any confirmation bias in participants performance in the *argument recall test*. Independent of the *debate display* condition, their performance for attitude-confirming arguments ($mean = 0.68, SE = 0.01$) was not significantly different to their performance for attitude-opposing arguments (*AO argument recall:* $mean = 0.64, SE = 0.02$).

Interaction with elements of the debate page Concerning clicks on the *show more* button, we observed that a high rate of participants in the debate display conditions with summary clicked on it (82.5%). We further observed that those participants who clicked on the button performed on average better in the argument recall test ($mean = 0.70, SE = 0.013$) than the participants who did not click on it ($mean = 0.59, SE = 0.038$). When comparing the three *debate display* conditions with summary, we observed only minor differences in the proportion of participants who clicked on the *show more* button (neutral suggestion: 76%, personalized persuasive suggestion: 88%, random persuasive suggestion: 84%). Regarding whether participants made contributions to the debate, we observed that across all *debate display* conditions, 81% of all participants made a contribution. Comparing the proportion of participants who made a contribution to the debate between the four *debate display* conditions, we observed only minor differences (no summary: 81%, neutral suggestion: 73%, personalized persuasive suggestion: 80%, random persuasive suggestion: 90%). Concerning the time that participants spent on the debate page, we observed that participants who saw summaries with neutral suggestions spent on average slightly less time (reported in seconds) on the debate page ($mean = 198, SE = 20$) than participants in all other *debate display* conditions ($mean = 232, SE = 11$). Further, participants who clicked the *show more* button spent on average more time on the debate page ($mean = 230, SE = 12$) than participants who did not click it ($mean = 178, SE = 25$).

Attitude change Attitude change was calculated as the difference of attitudes reported in the pre-interaction and the post-interaction questionnaire on a seven-point Likert scale ranging from "strongly disagree"(0) to "strongly agree"(6). Negative attitude change indicates a weakening of the initial attitude. In average, participants' attitude after engaging with the debate page was less extreme for all *debate display* conditions. The average attitude change was with -0.9 to -1.1 stronger for participants who saw a debate summary than with -0.6 for those who did not.

Detailed argument recall. When comparing the proportion for arguments that were displayed on the debate page (correct response: yes) and arguments that were not displayed on the debate page (correct response: no) we observed that a correct response was given for 82.5% of arguments that were displayed, while for arguments that were not displayed the proportion of correct responses was only at 42.6%.

5 DISCUSSION

Users need to be able to grasp the state of the debate, understand and take in what arguments were made and what viewpoints they transport to form well-informed attitudes without having to expend too much effort that might cause increased vulnerability to cognitive biases. With this objective in mind, we tested whether summaries and personalized persuasive suggestions to interact with the summary could affect participants' argument recall, specifically for attitude opposing arguments with a focus on confirmation bias mitigation.

We did not find evidence for an **effect of the summary** on attitude-opposing argument recall. However, in the exploratory data on participants' interaction with the *show more* button (only for debate display with summary), we observed that, on average, participants who clicked on the button performed better in the argument recall test. While this might imply that seeing all six arguments of the summary improved the performance in the argument recall test, this observation might also reveal that both, performance in the argument recall test and clicks on the show button, were affected by participants' enthusiasm for the task. This second explanation is supported by our observation that participants who clicked the *show more* button on average spent more time on the debate page than participants who did not click the button. Our exploratory observations of contributions to the debate across debate display conditions point towards no effect of summaries on participants' motivation to contribute to the debate. We did observe in the exploratory data on attitude change after being exposed to the debate page that participants who saw a summary demonstrated on average greater attitude change (attitude weakening) than participants who did not see a summary. If targeted follow-up studies confirmed this observation, this would indicate that summaries mitigate the confirmation and disconfirmation bias of reasoning found by Karlsen et al. [18]. We did not find evidence for an **effect of personalized persuasive suggestions** on attitude-opposing argument recall. However, this was somewhat expected since we did not find an effect of summaries overall, regardless of the suggestion. Against our assumption, we did not observe any difference in recall of attitude-confirming compared to attitude-opposing arguments, thus no **confirmation bias**.

5.1 Possible explanations for lack of findings and ideas for follow-up research

Not observing any effect of either the summary or the personalized persuasive suggestion might imply that these interventions do not facilitate argument recall and healthy engagement with online debates. An alternative explanation, however, could be that some properties of our experimental design prevented us from observing any effect of facilitation and confirmation bias mitigation: Our setup might have failed to sufficiently reflect the cognitively demanding, highly stimulating environment with endless possible choices of what to engage with in a limited amount of time that users are confronted with during real web interactions [21]. First, we infer from participants' positive feedback and high engagement that most participants were unexpectedly invested in the task.

Second, the contributions to the debate sampled from the *IBM ARGKP dataset* had to fulfill extensive quality control measures and thus did not reflect the low-quality contributions found in real online debates. Third, we only displayed 18 contributions, of which three expressed the same argument. Lastly, participants could explore the debate page as long as they wanted. If participants did not experience high cognitive demand that made them vulnerable to cognitive biases, they did not need a summary, let alone persuasive suggestions to interact with the summary, to understand the arguments made in the debate. This explanation would lead us to expect a ceiling effect of many participants who achieved 100% correct answers in the *argument recall* test. However, the mean score across all conditions of *debate display* was at 66.5% correct answers which is slightly above the proportion of arguments displayed in the debates (six out of ten). When comparing the distribution of responses for arguments displayed to those not displayed, we can observe a higher proportion of accurate responses for the former while participants struggled to categorize not displayed arguments correctly. Thus, the task of correctly categorizing not displayed arguments in the debate was difficult across conditions and the 66.5% of accurate responses might indeed represent a ceiling effect for accurate *yes* responses across debate display conditions.

We propose the following **alterations to the experimental design** to better assimilate real online debates for future user studies on approaches to boost healthy engagement with online debates: (1) Increasing the number of contributions and expressed arguments; (2) Including low-quality contributions; (3) Allowing for more interaction with the debate page (e.g., selecting between multiple debates on similar topics); (4) Introducing some distraction or time pressure to the scenario; (5) Adaption of the dependent variable that measures confirmation bias during argument recall by only calculating the proportion of correct *yes* responses for arguments that were displayed in the debate.

Further, we propose the following **new hypotheses**: (1) Debate summaries facilitate grasping the state of an online debate and enable users to accurately recall arguments made in the debate *(Reasoning: In the exploratory data, we observed a higher average of correctly recalled arguments that were presented in the debate for all participants that saw a summary. We further observed that participants who saw a summary with a neutral suggestion spent on average less time on the debate page while still performing equally well in the argument recall test as participants in other conditions)*; (2) Debate summaries mitigate attitude reinforcement due to confirmation bias and disconfirmation bias when processing arguments *(Reasoning: In the exploratory data, we observed that the average attitude weakening is greater for participants who were presented with a debate summary. We thus expect that summaries might support users to better process attitude-opposing arguments than they were found to be processed in debates without summaries by Karlsen et al. [18]).*

5.2 Ethical considerations

The approach of boosting online debate with summaries and personalized persuasive suggestions raises several ethical concerns that we will discuss in the following.

While **nudging** is supposedly done for the good of a person, the objective of influencing someone's decision making raises concerns about a paternalistic conception of users. Caraban et al. [4]

and Hansen and Jespersen [13] argue that while nudges that manipulate behavior raise ethical concerns, others do not. In the authors' taxonomy, the ones that do not are *transparent* regarding their objective and attempt to *prompt reflective choice*. We consider the summaries and suggestions that we applied to fall into this less concerning category. The **summaries** we displayed to the participants represented two viewpoints, supporting and opposing. This does neither capture and reflect different strengths of viewpoints nor debates with more dimensions than one supporting and one opposing viewpoint. On a real-world debate page minority viewpoints might be neglected in the summary, raising issues of fairness. This issue gets further pronounced by the vulnerability of online debates to malicious intentions. For example, users might propagate a certain viewpoint that is not actually shared by many others, or that is based on false information to become a majority viewpoint in the debate. This would distort the actual distribution of attitudes on the topic, potentially causing additional cognitive biases (such as *conformity bias and false consensus effect*). *Automation bias*, the tendency to over-rely on system-generated output, might be another problem of nudging with summaries [28]: users might trust them and judge them to be more accurate than they actually are. Following these concerns, providers of a debate page should be transparent about the summary generation and on its potential shortcomings. Further ethical concerns are risen by **personalized nudging** which, while intended to lead to beneficial outcomes for the user, could also be harmful. First, generating a user model or profile requires knowledge of user data, the collection and storage of which can compromise users' privacy. Second, personalization might diminish user control and the objective of it might not align with users' objectives. Real-world applications of personalized nudging should require transparency about the collected data, users' active consent, and give users control over their profile and experience.

5.3 Limitations

While the *IBM ArgKP dataset* offered the contributions and keypoints we required for this user study, we were restricted by the bounds of the data set. First, we could only consider eight out of 28 topics in the dataset for our user study, since those were the only topics with ten or more keypoints, required for the summary and *argument recall test*. In our pretests, we found that participants' attitude on most of these topics was rather imbalanced (unequal number of participants supporting and opposing the topic statement). However, we attempted to mitigate confounds due to this imbalance by introducing a quota of participants with the same attitude and stance on a topic and did not admit any further participants once this quota was reached. Further, it is likely that many of our participants are living in countries in which the three selected topics are not applicable (e.g., from countries that do not have capital punishment). Although we asked our participants to think about the topics from a worldwide perspective, we recognize that participants who are not affected by a topic would interact differently with debates on it. Lastly, the dataset represented viewpoints of contributions and keypoints merely as either supporting or opposing, which is a severe simplification of viewpoints (for considerations on more nuanced and comprehensive viewpoint labels see Draws et al. [7]). Another limitation of this study is that we only observed a single session of interaction with the debate page on one topic, preventing observations of potential long-term effects. Finally, while we attempted to ensure good quality of the persuasive suggestions with the mapping pretest, we realized that it is challenging to come up with suggestions for certain persuasion principles that do not overlap with other persuasion principles (i.e., reciprocity, liking, and consensus). Additionally, some of the suggestions might have been too long or too paternalistic, which were attributes that we did not test for in the pretest.

6 CONCLUSION

With the objective to boost online debates and overcome their downsides, such as confirmation bias, we conducted a user study to test the effects of summaries and personalized persuasive suggestions to engage with the summary on participants' attitude-opposing argument recall. We did not find evidence for an effect of either the summary or the personalized persuasive suggestions and did not observe confirmation bias for argument recall. However, additionally collected exploratory data motivated us to question whether no effects exist or whether properties of the study design prevented us from finding evidence for existing effects. This question remains to be answered in follow-up studies, for which we propose improvements of the study design. Motivated by exploratory data, we additionally formulate two new hypotheses on the effect of debate summaries.

ACKNOWLEDGMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860621.

REFERENCES

[1] Lucas Anastasiou and Anna De Liddo. 2021. Making Sense of Online Discussions: Can Automated Reports Help? In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.

[2] Ashley A. Anderson, Dominique Brossard, Dietram A. Scheufele, Michael A. Xenos, and Peter Ladwig. 2014. The "Nasty Effect:" Online Incivility and Risk Perceptions of Emerging Technologies*. *Journal of Computer-Mediated Communication* 19 (April 2014), 373–387. https://doi.org/10.1111/jcc4.12009

[3] Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From Arguments to Key Points: Towards Automatic Argument Summarization. *arXiv:2005.01619 [cs]* (June 2020). arXiv:2005.01619 [cs]

[4] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15.

[5] Robert B. Cialdini and Robert B. Cialdini. 2007. *Influence: The Psychology of Persuasion*. Vol. 55. Collins New York.

[6] Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64 (2014), 658–679. https://doi.org/10.1111/jcom.12104

[7] Tim Draws, Oana Inel, Nava Tintarev, Christian Baden, and Benjamin Timmermans. 2022. Comprehensive Viewpoint Representations for a Deeper Understanding of User Interactions With Debated Topics. In *Proceedings of the 2022 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '22)*. ACM, New York, NY, USA, 11. https://doi.org/10.1145/3498366.3505812

[8] Katharina Esau, Dannica Fleuß, and Sarah-Michelle Nienhaus. 2021. Different Arenas, Different Deliberative Quality? Using a Systemic Framework to Evaluate Online Deliberation on Immigration Policy in Germany. *Policy & Internet* 13 (2021), 86–112. https://doi.org/10.1002/poi3.232

[9] Katharina Esau, Dennis Friess, and Christiane Eilders. 2017. Design Matters! An Empirical Analysis of Online Deliberation on Different News Platforms. *Policy & Internet* 9 (2017), 321–342. https://doi.org/10.1002/poi3.154

[10] Jonathan St BT Evans. 2008. Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annu. Rev. Psychol.* 59 (2008), 255–278.

[11] L. Goette, H. J. Han, and B. T. K. Leung. 2020. *Information Overload and Confirmation Bias.* Working Paper. Faculty of Economics, University of Cambridge. https://doi.org/10.17863/CAM.52487

[12] Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (April 2020), 7805–7813. https://doi.org/10.1609/aaai.v34i05.6285

[13] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4, 1 (2013), 3–28.

[14] Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science* 12 (Nov. 2017), 973–986. https://doi.org/10.1177/1745691617702496

[15] Thomas T Hills. 2019. The Dark Side of Information Proliferation. *Perspectives on Psychological Science* 14 (2019), 323–330.

[16] Maurits Kaptein, Boris De Ruyter, Panos Markopoulos, and Emile Aarts. 2012. Adaptive Persuasive Systems: A Study of Tailored Persuasive Text Messages to Reduce Snacking. *ACM Transactions on Interactive Intelligent Systems* 2 (June 2012), 10:1–10:25. https://doi.org/10.1145/2209310.2209313

[17] Maurits Kaptein, Panos Markopoulos, Boris de Ruyter, and Emile Aarts. 2015. Personalizing Persuasive Technologies: Explicit and Implicit Personalization Using Persuasion Profiles. *International Journal of Human-Computer Studies* 77 (May 2015), 38–51. https://doi.org/10.1016/j.ijhcs.2015.01.004

[18] Rune Karlsen, Kari Steen-Johnsen, Dag Wollebæk, and Bernard Enjolras. 2017. Echo Chamber and Trench Warfare Dynamics in Online Debates. *European Journal of Communication* 32 (June 2017), 257–273. https://doi.org/10.1177/0267323117695734

[19] Scott O Lilienfeld, Rachel Ammirati, and Kristin Landfield. 2009. Giving Debiasing Away: Can Psychological Research on Correcting Cognitive Errors Promote Human Welfare? *Perspectives on psychological science* 4 (2009), 390–398.

[20] Philipp Lorenz-Spreen, Michael Geers, Thorsten Pachur, Ralph Hertwig, Stephan Lewandowsky, and Stefan M. Herzog. 2021. Boosting People's Ability to Detect Microtargeted Advertising. *Scientific Reports* 11 (July 2021), 15541. https://doi.org/10.1038/s41598-021-94796-z

[21] Philipp Lorenz-Spreen, Stephan Lewandowsky, Cass R. Sunstein, and Ralph Hertwig. 2020. How Behavioural Sciences Can Promote Truth, Autonomy and Democratic Discourse Online. *Nature Human Behaviour* 4 (Nov. 2020), 1102–1109. https://doi.org/10.1038/s41562-020-0889-7

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

[23] Gordon Pennycook and David G. Rand. 2019. Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning than by Motivated Reasoning. *Cognition* 188 (July 2019), 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

[24] Gloria Phillips-Wren and Monica Adya. 2020. Decision Making under Stress: The Role of Information Overload, Time Pressure, Complexity, and Uncertainty. *Journal of Decision Systems* 29 (Aug. 2020), 213–225. https://doi.org/10.1080/12460125.2020.1768680

[25] Emmanuel M. Pothos, Stephan Lewandowsky, Irina Basieva, Albert Barque-Duran, Katy Tapper, and Andrei Khrennikov. 2021. Information Overload for (Bounded) Rational Agents. *Proceedings of the Royal Society B: Biological Sciences* 288 (Feb. 2021), 20202957. https://doi.org/10.1098/rspb.2020.2957

[26] Alisa Rieger, Tim Draws, Mariët Theune, and Nava Tintarev. 2021. This Item Might Reinforce Your Opinion: Obfuscation and Labeling of Search Results to Mitigate Confirmation Bias. In *Proceedings of the 32st ACM Conference on Hypertext and Social Media.* ACM, Virtual Event USA, 189–199. https://doi.org/10.1145/3465336.3475101

[27] Jack B Soll, Katherine L Milkman, and John W Payne. 2015. A User's Guide to Debiasing. *The Wiley Blackwell handbook of judgment and decision making* 2 (2015), 924–951.

[28] Stefan Strauß. 2021. Deep Automation Bias: How to Tackle a Wicked Problem of AI? *Big Data and Cognitive Computing* 5 (June 2021), 18. https://doi.org/10.3390/bdcc5020018

[29] Charles S. Taber, Damon Cann, and Simona Kucsova. 2009. The Motivated Processing of Political Arguments. *Political Behavior* 31 (June 2009), 137–155. https://doi.org/10.1007/s11109-008-9075-8

[30] Charles S. Taber and Milton Lodge. 2006. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science* 50 (2006), 755–769. https://doi.org/10.1111/j.1540-5907.2006.00214.x

[31] Ben M. Tappin, Leslie van der Leer, and Ryan T. McKay. 2017. The Heart Trumps the Head: Desirability Bias in Political Belief Revision. *Journal of Experimental Psychology. General* 146 (Aug. 2017), 1143–1149. https://doi.org/10.1037/xge0000298

[32] Richard H Thaler and Cass R Sunstein. 2008. Nudge: improving decisions about health. *Wealth, and Happiness* 6 (2008), 14–38.

[33] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185 (Sept. 1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124