Interventions for Softening Can Lead to Hardening of Opinions: Evidence from a Randomized Controlled Trial

Andreas Spitz*
EPFL
Switzerland
andreas.spitz@epfl.ch

Ahmad Abu-Akel*
University of Lausanne
Switzerland
ahmad.abuakel@unil.ch

Robert West
EPFL
Switzerland
robert.west@epfl.ch

ABSTRACT

Motivated by the goal of designing interventions for softening polarized opinions on the Web, and building on results from psychology, we hypothesized that people would be moved more easily towards opposing opinions when the latter were voiced by a celebrity they like, rather than by a celebrity they dislike. We tested this hypothesis in a survey-based randomized controlled trial in which we exposed respondents to opinions that were randomly assigned to one of four spokespersons each: a disagreeing but liked celebrity, a disagreeing and disliked celebrity, a disagreeing expert, and an agreeing but disliked celebrity. After the treatment, we measured changes in the respondents' opinions, empathy towards the spokespersons, and use of affective language.

Unlike hypothesized, no softening of opinions was observed regardless of the respondents' attitudes towards the celebrity. Instead, we found strong evidence of a hardening of pre-treatment opinions when a disagreeing opinion was attributed to an expert or when an agreeing opinion was attributed to a disliked celebrity. We also observed a pronounced reduction in empathy for disagreeing spokespersons, indicating a punitive response. The only celebrity for whom, on average, empathy remained unchanged was the one who agreed, even though they were disliked.

Our results could be explained as a reaction to violated expectations towards experts and as a perceived breach of trust by liked celebrities. They confirm that naïve strategies at mediation may not yield intended results, and how difficult it is to depolarize—and how easy it is to further polarize or provoke emotional responses.

ACM Reference Format:
Andreas Spitz, Ahmad Abu-Akel, and Robert West. 2021. Interventions for Softening Can Lead to Hardening of Opinions: Evidence from a Randomized Controlled Trial. In *Proceedings of The Web Conference 2021 (WWW '21), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442381.3450019

1 INTRODUCTION

Soaring polarization, and a concomitant decrease in civility, pose a major challenge to today's society. Research shows that in Europe [18, 38, 49] as well as in the United States [3, 5, 35], trust in the mainstream media, public trust in science, and mutual esteem between people of different political convictions have plummeted drastically over the last two decades, while partisanship among politicians has starkly increased [1, 21]. People have always disagreed with each other. However, the current political climate is rendered fundamentally without precedence by the existence of the Web, where extreme opinions, biased information, and fabricated false facts ("fake news") are proliferating and amplifying in echo chambers [54] and filter bubbles [16].

To address these alarming trends, fight polarization, and ultimately make the Web a more civil place, there is a need to convince Web users to be respectful and seriously consider divergent opinions. Social computing researchers have suggested to do so by explicitly exposing users to posts from across the aisle [20, 33, 41]. Other research has, however, shown that these interventions tend to not be effective, possibly due to confirmation bias [42], where people tend to quickly dismiss divergent opinions when they are too far removed from their own [4]. This outcome is in line with psychological research, where it is well established that people maintain confirmation bias in order to avoid cognitive dissonance and readily accept opinions that support their own, while dismissing arguments from the other side [42]. Thus, haphazardly exposing people to opinions from the other side of the aisle even runs the risk of making things worse, by inciting spite and decreasing trust in the mainstream media [3].

Recent research in psychology has addressed this issue by evaluating nuanced strategies for nudging people with extreme opinions into revising them, which shows that a softening of extreme opinions is possible—if done carefully. For instance, Bruneau and colleagues [9, 10] demonstrated that collective blame of Muslims for individual acts of violence, and thereby anti-Muslim hostility, could be reduced by highlighting hypocrisy (e.g., by making a Christian who collectively blamed Muslims realize that they would not collectively blame Christians for acts of violence committed by individual Christians). Similarly, Hameiri and colleagues [23, 24] have demonstrated the effectiveness of paradoxical thinking interventions for softening extreme opinions.

Inspired by these works (see also the related work in Section 2) and with the eventual goal of making the Web a better place, we initially set out to explore further strategies for softening polarized opinions. Psychology has shown that people frequently identify with celebrities, building a pseudo-personal, one-way rapport with them [6], a phenomenon that has for a long time been leveraged by marketeers, who successfully use celebrities as spokespersons in product ads [17, 39]. Hence, it seems reasonable to suspect that people might be more willing to accept an opposing opinion if it comes from a celebrity they know, like, and trust. Indeed, there is strong evidence to suggest that liked persons [50], including

*Both authors contributed equally to this work.

WWW '21, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8312-7/21/04.
https://doi.org/10.1145/3442381.3450019

celebrities [7, 27, 28], can shift others' opinions, and that individuals tend to adapt their opinions and attitudes to persons whom they like and with whom they identify [14, 26].

Building on these results, we hypothesized that

(1) people would be more easily convinced to soften their extreme opinions if the opposite opinion was voiced by a celebrity they liked, compared to a celebrity they disliked. Additionally, we hypothesized that
(2) people would be more easily convinced by an expert towards whom they had no prior disposition than by a disliked celebrity, and that
(3) people could be pushed away from their current opinion if that opinion was shared by a celebrity they disliked.

Testing the above hypotheses is not straightforward, since people's opinions about topics and their sympathies for others are intricately intertwined, due to homophily [37] (we tend to form bonds with those who are like us) and social influence [53] (we tend to become like those with whom we already share bonds). For these reasons, it can be difficult to find, for a given person and a given topic, a celebrity whom the person likes but with whom they disagree on the topic.

To exert finer control and to systematically tease apart agreement from liking, we designed a survey-based randomized controlled trial (RCT) that allowed us to attribute any opinion to any spokesperson, and thus to measure the causal effect of a respondent's fondness of a spokesperson on the spokesperson's ability to depolarize the respondent's opinions. In particular, we tested the effects of attributing disagreeing opinions to liked celebrities vs. disliked celebrities vs. experts, as well as the effect of attributing agreeing opinions to disliked celebrities. Additionally, we measured how being exposed to a spokesperson's opinion shifted the respondent's empathy for that spokesperson.

Results obtained by deploying the study to 379 respondents on the Amazon Mechanical Turk crowdsourcing platform refuted our initial hypotheses. No matter whether the respondent liked or disliked the celebrity to whom an opposing opinion was attributed, we, on average, never observed a softening of opinions. Neither did neutral experts provide the expected result. Opposite to our hypothesis, we found strong evidence of further hardening of opinions in the expert condition (i.e., away from the expert's opinion, further polarizing the respondent's previously held opinion). We also found evidence of vindictive behavior, whereby disagreement by a spokesperson was "punished" by a reduction in the respondent's empathy for pain that the spokesperson experienced. The only spokesperson for whom empathy remained unchanged was the one who agreed with the respondent, despite being disliked. Analyzing linguistic cues in written opinion statements provided by the respondents supported these findings.

These results confirm that naïve strategies for softening extreme opinions may yield counterproductive results, and highlight how hard it is to depolarize opinions—and how easy it is to create emotional reactions and further polarize already-polarized opinions.

Contributions. To summarize our contributions, we

(1) designed a survey-based RCT to measure the aptitude of celebrity and expert spokespersons for softening polarized opinions (Section 3);
(2) deployed the survey to 379 carefully selected respondents on Amazon Mechanical Turk (Section 4); and
(3) showed that in no condition was depolarization achieved. The only observed shift was a further polarization toward previously held opinions (Section 5). We conclude the paper by discussing the implications of these results (Section 6).

2 RELATED WORK

In addition to the above-cited works, our research is informed by work that can be categorized into three groups, namely *polarization and the backfire effect*, the role of *celebrity spokespersons*, and the relevance of *conveying scientific agreement*.

Polarization and backfire effect. The difficulty of changing people's previously held opinions and attitudes is well known [25]. While personal engagement has been shown to reduce polarization even on partisan topics in offline settings [19], depolarization in online settings appears to be especially challenging. Attempts at depolarizing extreme opinions held by Web users may lead to undesired and even opposite effects, commonly known as the "backfire effect" [43]. For example, studies on the exposure to opposing views on social media have demonstrated a resulting increase, rather than a decrease, in political polarization [4]. Even actively aiming to increase empathy by creating situations in which people are encouraged to "put themselves in the shoes of others" and to take another's perspective holds the potential for creating such a backfire effect [12]. However, the existence of the effect is debated in the literature, and other studies have found no evidence for it [58]. Various explanations have been offered for this discrepancy in findings, including the observation that the susceptibility to polarizing opinions may be explained by a lack of reasoning, rather than by motivated reasoning [46], which might entirely preclude a change in opinion, depending on the situation in which the observation of a change is expected. We extend this line of research by exposing respondents to conditions that vary in terms of the identity of the spokesperson (celebrity vs. scientific/academic expert), the attitude towards the spokesperson (like vs. dislike), and opinion congruence (agree vs. disagree).

Celebrity spokespersons. The use of celebrities as spokespersons is well researched in marketing [17, 39], showing that the one-sided, pseudo-personal bond that everyday people form with celebrities [6] can be exploited to influence their behavior. While the similar effect of celebrity endorsement of political ideas has been considered in prior work [27], the effect of a celebrity spokesperson on the level of polarization is not clearly established. Celebrities with whom the public is merely familiar but towards whom they do not have a favorable opinion may have a negative effect, while celebrities who are viewed favorably consistently have positive effects [27]. In this study, we experimentally investigate the effect of celebrity spokesperson attribution on polarization by exposing respondents to *agreeing* or *disagreeing* opinions of celebrities whom they *like* or *dislike*.

Conveying scientific agreement. Misinformation in general and conspiracy theories in particular are known to have harmful consequences. Indeed, the belief in such theories has been linked

to serious societal problems, such as vaccine hesitancy [29], climate change denial [31], extremist political views [56], and prejudice [30, 32]. It is clear that scientific results play a role in opinion formation, and it has been shown that the communication of scientific agreement can neutralize the politicization of facts [55] and thus counteract polarization. However, other research shows that there are tendencies in partisan environments to deny the correct interpretation of scientific results when that interpretation conflicts with previously held opinions [57]. Therefore, in this study, we also investigate the effect of a scientific/academic expert spokesperson on the softening of polarization, and how this effect compares to that of celebrities.

3 STUDY DESIGN

In an effort to mellow extreme opinions gently, we investigated if the attribution of opposing views to well-liked celebrities would have a softening effect on extreme viewpoints that people held on a given topic. To measure the hypothesized effect, we designed a survey-based randomized controlled trial (RCT) in which any opinion could be attributed to any spokesperson. Randomizing the opinion-to-spokesperson attribution thus allowed us to elicit the causal effects of the different types of spokesperson on respondents' opinions. Our design is summarized schematically in Fig. 1 and briefly described here, whereas further details on the individual components are given in the following subsections.

As motivated in the introduction, we tested four conditions:

- disagree/like: **disagreeing** opinion, **liked** celebrity
- disagree/dislike: **disagreeing** opinion, **disliked** celebrity
- disagree/expert: **disagreeing** opinion, **expert** spokesperson
- agree/dislike: **agreeing** opinion, **disliked** celebrity

As for the topics on which opinions were expressed, we chose four contemporary and controversial topics of societal interest (Section 3.1).

When compiling a pool of potential celebrity spokespersons, we took care to ensure that the selected celebrities' actual opinions on the topics were likely to be unknown to the respondents in our study (cf. Appendix A.1), as known opinions would have compromised the validity of randomizing quotations to spokespersons.

For each topic, we selected two real quotations from a large news corpus, one on either side of the polarization spectrum (cf. Appendix A.2). The quotations were generic enough such that they could be attributed to any spokesperson from our pool.

Selecting an opinion and a spokesperson for a given respondent, topic, and condition required knowing the respondent's opinion about the topic and their attitude towards the spokesperson. We therefore elicited this information in a screening survey (Section 3.2) that we conducted before the actual RCT.

As a result of the above steps, we obtained, for each respondent and each of the four topics, a liked and a disliked celebrity, as well as an agreeing and a disagreeing opinion, such that each experimental condition could be instantiated with a concrete opinion–spokesperson pair. For the expert condition, we invented a fictitious person named "Dr. Michael Barnes" who could be cited as an expert on all topics. Using an invented expert further ensured that no respondent could have a prior attitude towards the expert.

For each topic, respondents were randomly and uniformly assigned to one of the conditions (Section 3.3). These random assignments were then used for the main survey (Section 3.4), which consisted of three steps:

(1) Measuring the respondent's **pre-treatment** opinion about the topic and their pre-treatment empathy towards the spokesperson that had been randomly assigned to the topic.
(2) Administering the **treatment:** exposing the respondent to the opinion attributed to the spokesperson, via a mock task in which it was stated that the spokesperson had uttered the opinion. The respondent was then asked to summarize the spokesperson's opinion and describe how clear and convincing they found it.
(3) Measuring the respondent's **post-treatment** opinion about the topic and their post-treatment empathy towards the spokesperson, by repeating the pre-treatment questions.

In the remainder of this section, we provide more details on the individual aspects of the RCT.

3.1 Choice of Topics

Topic selection was motivated by two factors: (*i*) established opinion surveys that could be adapted for our purposes had to exist in the literature on the topic, and (*ii*) the respondent's opinion on the topic could be expressed on a linear scale with opposing viewpoints at each end. Furthermore, since any temporally limited interaction with written viewpoints is likely to result in only small changes to the respondent's opinions (i.e., a small effect size should be expected), we also required topics with extreme polar viewpoints, with which respondents were likely to be familiar. To this end, we selected four contemporary topics of current societal interest, for which we considered viewpoints at the opposing ends of the polarity spectrum:

(1) **Climate change:** "Climate change is a serious threat that all of us need to address" vs. "The risks of climate change are vastly exaggerated".
(2) **Vaccination:** "Vaccines are not harmful, they save lives and protect the community" vs. "Many vaccines have serious side effects and can cause severe illness".
(3) **Immigration:** "Immigrants take jobs away from Americans and undermine our culture" vs. "Immigrants are good for our economy and immigration benefits all of us".
(4) **Abortion:** "Abortions should not be allowed for anyone under any circumstances" vs. "Abortion should be allowed and should be the choice of prospective parents alone".

3.2 Screening Survey

In preparation of the survey that contained the RCT, we conducted a screening survey to identify viable respondents and establish their opinions and their attitudes towards the pool of possible spokespersons. The screening survey established the respondents' opinion on each of the topics on a 7-point Likert scale, and the attitude towards all candidate celebrity spokespersons to determine whether they liked, disliked, did not know, or were ambivalent towards the respective spokesperson. We also used this screening survey to collect basic demographic data, including the state of residence in the US, age, gender, and level of education. In the RCT,

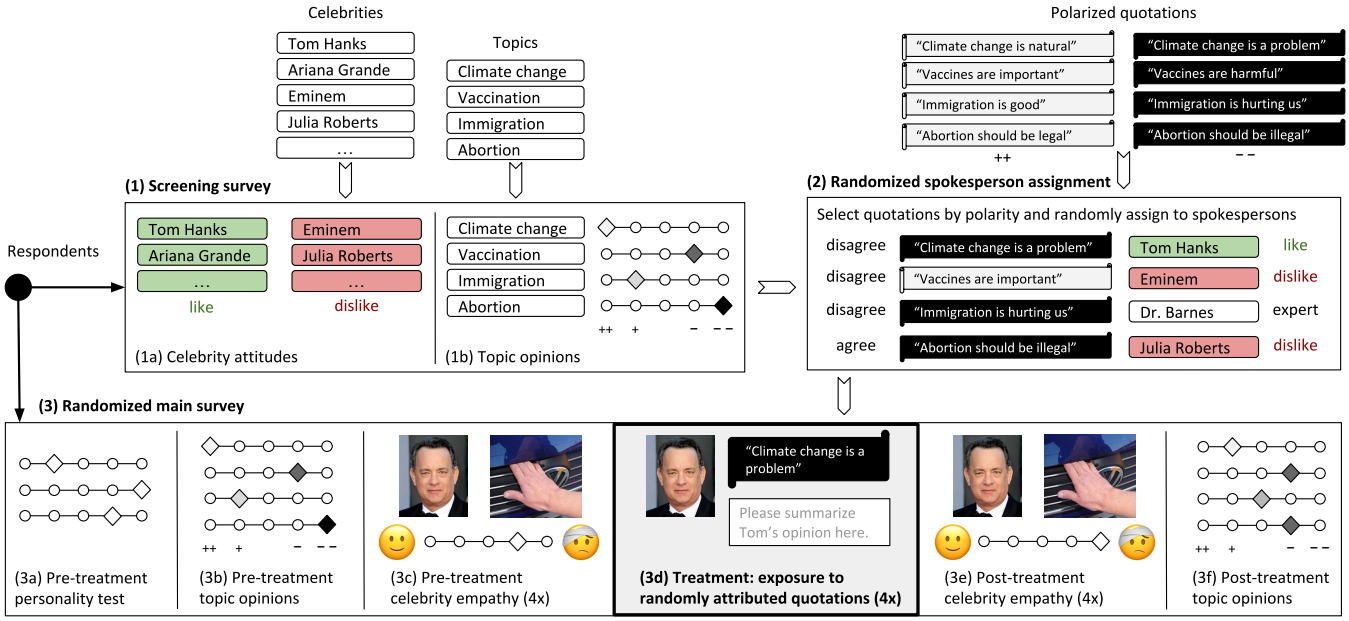

Figure 1: Overview of the randomized controlled trial. (1) Respondents' attitudes towards celebrities and opinions on topics are elicited in a screening survey. (2) Respondents are randomly assigned one of 4 experimental conditions per topic. Spokesperson–quotation pairs that match the conditions are generated based on the screening responses. (3) The main survey establishes respondents' personality, as well as their opinions on the 4 topics and their empathy for the 4 spokespersons in an empathy-for-pain test before and after treatment. Treatment consists of 4 separate interactions with the assigned quotations (one per topic) that are attributed to the assigned spokespersons. (Pain test image adapted from Shamay-Tsoory et al. [48].)

respondents were assigned spokespersons in personalized surveys, based on their responses to the screening survey.

3.3 Randomized Spokesperson Assignment

For each topic, respondents were randomly and uniformly assigned to one of the 4 conditions listed at the beginning of Section 3, such that each topic and each condition occurred once per respondent.

The specific quotation that was shown to a respondent was determined by the randomly assigned condition and her opinion on the topic as previously stated in the screening survey. For the disagreeing conditions, we selected the quotation at the opposite end of the polarity spectrum (or a random quotation if the respondent had stated to have a neutral opinion in the screening survey). For the agreeing condition, the matching quotation was selected.

Spokespersons were randomly assigned in a similar fashion according to the like or dislike condition from the pool of celebrities that were indicated to be liked or disliked by the respondent during the screening survey. The attribution of quotations to spokespersons served solely to gauge the respondents' reaction to this attribution, and none of the quotations in the study were actually uttered by any of the attributed spokespersons (all used quotations and their actual sources are listed in Appendix C).

3.4 Randomized Main Survey

The main survey containing the RCT consisted of 3 phases: (1) pre-treatment, (2) treatment, and (3) post-treatment, each described in turn next. The survey took approximately 45 minutes to complete.

Pre-treatment opinions and empathy. At the beginning of the pre-treatment phase, we established respondents' *personality* via the Ten Item Personality Inventory (TIPI) [22].

We then gauged respondents' *opinions on each of the four topics* with a long questionnaire that we adapted from existing validated surveys on climate change [13], vaccination [15], immigration [47], and abortion [52]. Next, we asked respondents to summarize their opinion on the topics in 2–3 sentences (250 characters or more), and collected their condensed overall opinion on the topics on a 7-point Likert scale.

Furthermore, we introduced the respondents to the four specific spokespersons that were assigned to them for the treatment phase and collected their empathic responses towards the spokespersons with an *empathy-for-pain test* [28]: for each spokesperson, respondents were shown a portrait photograph, alongside one of four different images of a human hand in a painful situation that is likely to lead to an injury (cf. Fig. 1), in order to elicit an empathy response. The respondents were then asked to rate how much pain the spokesperson was likely to feel on a 7-point Likert scale. We also asked the respondents to state if they believed that they would be friends with the spokesperson, and tasked them to write a statement of 2–3 sentences (250 characters or more) expressing their attitude towards the spokesperson.

Treatment: exposure to randomly attributed quotations. The treatment phase was realized as a mock task with personalized assignments that depended on the respondent's responses in the

Table 1: Summary statistics of RCT respondents.

Age	Count	Percentage
Under 21	1	0.3%
21–30	63	17%
31–40	133	35%
41–50	92	24%
51–60	65	17%
Over 60	25	7%
Education		
No degree	5	1%
High-school graduate	123	33%
Bachelor's degree	187	49%
Graduate degree	64	17%
Gender		
Female	211	56%
Male	166	44%
Other	2	0.5%
Total	$N = 379$	100%

screening survey. For each of the four topics, respondents were shown a quotation on this topic alongside a portrait photograph of the spokesperson to which the quotation was attributed. Respondents were tasked to summarize the spokesperson's view and state their own opinion on how effective the spokesperson was at communicating their opinion: "Please summarize [spokesperson]'s position on [topic], and describe how clear and convincing you find the argument in 3–4 sentences (350 characters or more)".

Post-treatment opinions and empathy. In the post-treatment phase, the empathy-for-pain tests were repeated for each of the four spokespersons, followed by a repetition of the topic opinion questionnaires. These questionnaires were identical to the pre-treatment phase. Finally, respondents were asked to explicitly state their level of agreement with each of the four quotations on a 7-point Likert scale. After the survey, respondents were debriefed on the study and informed of the misattribution of the quotations.

4 SURVEY DATA

The study was conducted between August and November 2019. We recruited respondents for the surveys from the pool of crowdworkers on Amazon Mechanical Turk. We offered tasks only to workers with residence in the United States, an approval rate of previously performed tasks of at least 99%, and at least 5,000 previously approved tasks (for details about the study deployment, see Appendix A.3). We collected data from 379 respondents, whose demographic statistics are summarized in Table 1. Based on their pre-treatment responses, we found the respondents to be largely biased towards liberal views (for details, see Appendix B).

From the total of 1516 completed surveys (each of the 379 respondents completed the survey for four opinion–spokesperson combinations), we excluded 30 outlier data points (1.98%), defined as 3 standard deviations from the mean.

In preparation for the computation of personality scores, reverse-scored TIPI responses were reversed. Similarly, responses to the detailed topic-specific questionnaires were reversed where necessary for consistency such that higher scores corresponded to more conservative views and lower scores to more liberal views.[1]

A respondent's change in opinion towards each of the four topics was measured as the opinion reported post-treatment, minus the opinion reported pre-treatment. Since the opinion of each individual respondent could change towards either polarity, we aligned the opinion change in relation to the quotation that was shown to the respondent, such that a positive value corresponded to a change towards the polarity of the quotation that was used in the treatment, while a negative value indicated a shift away from the extreme polarity of the quotation.

Analogously, a respondent's change in empathy towards the spokesperson was measured as the post-minus-pre difference in the responses given to the empathy-for-pain question. All scores were standardized individually (pre-treatment and post-treatment) prior to computing the changes in opinion and empathy.

5 RESULTS

In this section, we present the results of the RCT, with a focus on the changes in *opinion* towards the topic, *empathy* towards the spokesperson, and the *language* that was used by the respondents in their opinion statements. We discuss the findings and their implications in Section 6.

5.1 Change in Opinion and Empathy for Pain

First, to investigate the potential impact of demographic data and personality on the respondents' opinions and empathy-for-pain scores, we computed Spearman's correlations of age, sex, education, and the five subscales of the TIPI personality questionnaire (extraversion, agreeableness, consciousness, emotional stability, and openness) with the mean change in opinion and empathy. We found no significant correlations ($p > 0.05$ in all cases).

Second, we performed two separate repeated-measures analyses of variance (ANOVA) to examine if the respondents' change in opinion and empathy varied across the four topics, against a null hypothesis of no difference in the means of change. The results revealed no significant differences in opinion change ($F(3, 1134) = 0.317$, $p = 0.813$) or empathy change ($F(3, 1134) = 0.411$, $p = 0.745$) across all four topics (see Fig. 2a, c).

Based on the above, the effects of the spokesperson condition on opinion and empathy-for-pain change were examined using multivariate analysis of variance (MANOVA) independent of topic and without controlling for age, sex, education, or personality traits. We employ a MANOVA in order to protect the analysis from the inflation of the overall Type I error rate that would otherwise be produced by multiple univariate tests [8]. For a significant MANOVA, we separately report the univariate analysis-of-variance (ANOVA) tests on individual dependent variables, i.e., the effect of the spokesperson condition on opinion change and empathy change. For significant ANOVAs (rejecting the null hypothesis of no difference in the means across the four spokesperson conditions) we then report false discovery rate (FDR; q-value = 0.05) corrected pairwise comparisons of means for all spokesperson conditions. Importantly, we also report FDR-corrected one-sample t-tests against

[1]The data is available at https://github.com/epfl-dlab/SpokespersonAttribution

a null hypothesis of no change (difference from zero). The multivariate test was significant (Pillai's trace = 0.013, $F(6, 2962) = 3.25$, $p = 0.003$), with a small effect size (Cohen's $d = 0.17$, which denotes the mean difference divided by the pooled standard deviation). In the following, we therefore report the effect of the spokesperson on opinion change and empathy change separately.

Opinion change by spokesperson condition. The model for opinion change across the four spokesperson conditions was significant ($F(3, 1481) = 3.62$, $p = 0.013$), with a small effect size (Cohen's $d = 0.17$). In Fig. 2b, we show the standardized changes in respondents' opinion pooled over topics, split by the spokesperson condition (i.e., the treatment variable). In a direct comparison between spokesperson conditions, FDR-corrected pairwise comparisons revealed a significant difference only in the amount of change between the agree/dislike and the disagree/expert conditions (mean difference ± standard error: $MD = 0.241 \pm 0.074$ with an FDR corrected $p_{\text{corr}} = 0.007$) with a small effect size (Cohen's $d = 0.25$). The difference between the conditions agree/dislike and disagree/dislike was also small (Cohen's $d = 0.14$) and significant, but did not survive correction ($MD = -0.144 \pm 0.073$, $p = 0.0498$, but $p_{\text{corr}} = 0.149$). Importantly, FDR-corrected one-sample t-tests (difference from zero) showed that both the agree/dislike ($p = 0.016$, $p_{\text{corr}} = 0.040$, Cohen's $d = 0.25$) and the disagree/expert ($p = 0.020$, $p_{\text{corr}} = 0.040$, Cohen's $d = 0.24$) conditions resulted in a significant opinion change. On average, the agree/dislike condition led to an entrenchment of the respondents' pre-treatment opinions, while the disagree/expert condition caused the respondents to distance themselves from the expert's polarity even further.

Empathy change by spokesperson condition. The model for the degree of empathy change across the four spokesperson conditions (Fig. 2d) was significant ($F(3, 1481) = 2.85$, $p = 0.036$), with a small effect size (Cohen's $d = 0.16$). Pairwise comparisons between the four spokesperson conditions revealed a small but significant difference between the disagree/like and the agree/dislike conditions ($MD = 0.131 \pm 0.048$, $p = 0.007$, $p_{\text{corr}} = 0.042$, Cohen's $d = 0.21$). The difference between the agree/dislike and the disagree/expert conditions was also small (Cohen's $d = 0.16$) and significant, but did not survive correction ($MD = -0.106 \pm 0.048$, $p = 0.028$, $p_{\text{corr}} = 0.084$). Importantly, FDR-corrected one-sample t-tests (difference from zero) show that respondents attributed, with small to medium effect sizes, less empathy-for-pain to spokespersons in the disagree/like ($p < 0.001$, $p_{\text{corr}} < 0.001$, Cohen's $d = 0.44$) and disagree/expert ($p = 0.002$, $p_{\text{corr}} = 0.004$, Cohen's $d = 0.32$) conditions, and marginally less to the spokesperson in the disagree/dislike condition ($p = 0.040$, $p_{\text{corr}} = 0.053$, Cohen's $d = 0.21$), in the post- versus the pre-treatment phase.

It is important to highlight that in a direct comparison of mean empathy levels before and after treatment, there were no significant differences between spokesperson conditions in the empathy rating before treatment ($F(3, 1481) = 0.27$, $p = 0.845$) or after treatment ($F(3, 1481) = 1.11$, $p = 0.345$) across the four spokesperson conditions. This shows that the change in empathy is not due to spokesperson-specific variation in the pre- or post-treatment phase, but rather due to the treatment itself.

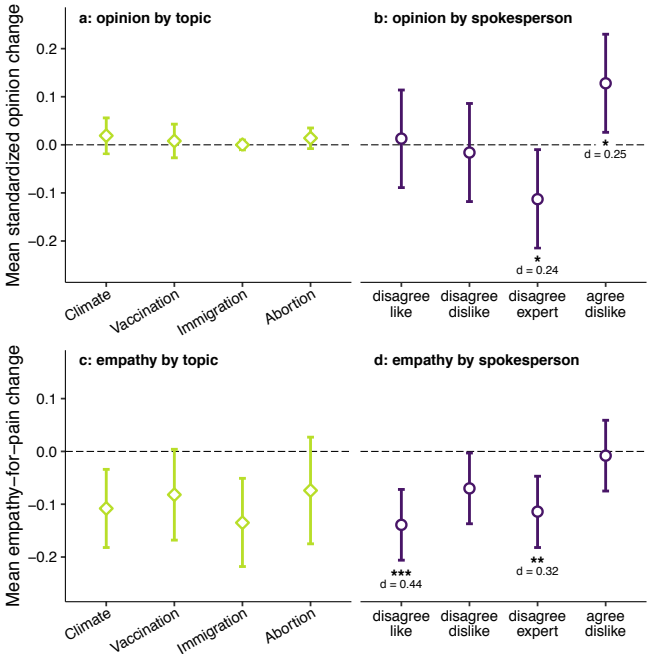

Figure 2: Mean change (post- minus pre-treatment) in opinion (a) by topic and (b) by spokesperson, and in empathy (c) by topic and (d) by spokesperson. Positive values denote a shift towards the treatment quotation or an increase in empathy, respectively. Asterisks denote significance levels of FDR-corrected p-values ($^{*}p_{\text{corr}} < 0.05$, $^{**}p_{\text{corr}} < 0.01$, $^{***}p_{\text{corr}} < 0.001$). Cohen's d is shown for significant effects.

5.2 Language Change

To analyze the potential change in sentiment that is contained in the statements in which respondents expressed their opinion towards the four topics before and after the treatment, we extracted words listed by Linguistic Inquiry and Word Count (LIWC) [51] in categories that capture positive or negative affective language. All LIWC word statistics are given as fractions of words in the statement that belong to the respective word classes: (*i*) negative affective words, (*ii*) positive affective words, and two subsets of the affective word class that consist of words conveying (*iii*) anger or (*iv*) sadness.

First, we computed Spearman's correlations to investigate the potential impact of demographic data and personality on the mean change in respondents' word usage frequency for each of the four topics. We found a few nominal significant correlations between the change in word frequency for positive and negative affective words, sadness, and anger with age, sex, and personality scores, but the number of these instances matched the expected number by chance for a 0.05 significance threshold (5 out of 128 correlations were significant), and none of these correlations survived correction at a Bonferroni significance threshold of 0.002). However, when we investigated the potential impact of topic on the mean change in respondents' word usage frequency, a repeated-measures ANOVA revealed significant differences between topics in the change of frequency of positive affective words ($F(3, 1134) = 4.27$, $p = 0.005$),

negative affective words ($F(3, 1134) = 7.035$, $p < 0.001$), and anger ($F(3, 1134) = 5.13$, $p = 0.002$). There were no significant differences between topics in the change of frequency of sadness-related word ($F(3, 1134) = 0.345$, $p = 0.793$).

Accordingly, we used a multivariate analysis of covariance (MANCOVA) to test the differences between spokesperson conditions in the frequency-of-use change for positive affective words, negative affective words, anger, and sadness, while controlling for variations across topics. The multivariate test was significant (Pillai's trace $= 0.024$, $F(12, 4437) = 2.99$, $p < 0.001$) with a small effect size (Cohen's $d = 0.18$). In the following, we therefore report the effect of the spokesperson condition on the change in the frequency of use of positive affective words, negative affective words, anger, and sadness, separately.

Positive affective language use. The model fitted to the change in respondents' usage of positive affective words (Fig. 3a) was significant ($F(3, 1480) = 4.12$, $p = 0.006$), with a small effect size (Cohen's $d = 0.18$). Pairwise comparisons between the four spokesperson conditions revealed a small but significant difference between the disagree/like and the agree/dislike conditions ($MD = 0.006 \pm 0.002$, $p = 0.001$, $p_{\text{corr}} = 0.006$, Cohen's $d = 0.25$). There was also a small and a significant difference between the disagree/expert and the agree/dislike conditions ($MD = 0.005 \pm 0.002$, $p = 0.012$, $p_{\text{corr}} = 0.036$, Cohen's $d = 0.21$). Importantly, FDR-corrected one-sample t-tests (difference from zero) showed that the change in the respondents' frequency of use was significant only in the agree/dislike spokesperson condition with a small to medium effect size ($p = 0.001$, $p_{\text{corr}} = 0.004$, Cohen's $d = 0.35$), where we observed an increase in positive affective word usage.

Negative affective language use. The model for the change in respondents' usage of negative affective words (Fig. 3b) was significant ($F(3, 1480) = 7.34$, $p < 0.001$), with a small effect size (Cohen's $d = 0.25$). Pairwise comparisons between the four spokesperson conditions revealed small but significant differences between the disagree/like condition and all other spokesperson conditions: agree/dislike ($MD = -0.007 \pm 0.002$, $p < 0.001$, $p_{\text{corr}} < 0.001$, Cohen's $d = 0.26$); disagree/dislike ($MD = -0.006 \pm 0.002$, $p < 0.001$, $p_{\text{corr}} = 0.001$, Cohen's $d = 0.26$); disagree/expert ($MD = -0.006 \pm 0.002$, $p < 0.001$, $p_{\text{corr}} = 0.001$, Cohen's $d = 0.25$). Importantly, FDR-corrected one-sample t-tests (difference from zero) showed that the change in respondents' frequency of use was significant only in the disagree/like spokesperson condition with a medium effect size ($p < 0.001$, $p_{\text{corr}} < 0.001$, Cohen's $d = 0.46$), where we observed an increase in negative affective word usage.

Indicators of anger and sadness. To further investigate the usage of negative affective language, we consider the subsets of negative affective words that indicate anger or sadness. In Fig. 3c, we show the results for words that indicate anger, for which the model was significant ($F(3, 1480) = 4.82$, $p = 0.002$), with a small effect size (Cohen's $d = 0.20$). Pairwise comparisons between the four spokesperson conditions revealed small but significant differences between the disagree/like and all the other spokesperson conditions: agree/dislike ($MD = -0.003 \pm 0.001$, $p < 0.001$, $p_{\text{corr}} = 0.002$, Cohen's $d = 0.23$); disagree/dislike ($MD = -0.002 \pm 0.001$, $p = 0.015$, $p_{\text{corr}} = 0.030$, Cohen's $d = 0.18$); disagree/expert ($MD = -0.002 \pm 0.001$, $p = 0.008$, $p_{\text{corr}} = 0.024$, Cohen's $d = 0.15$). Importantly, FDR-corrected one-sample t-tests (difference from zero) showed that the change in the respondents' frequency of use was significant only in the disagree/like condition ($p = 0.001$, $p_{\text{corr}} = 0.004$, Cohen's $d = 0.34$), where the level of anger increased.

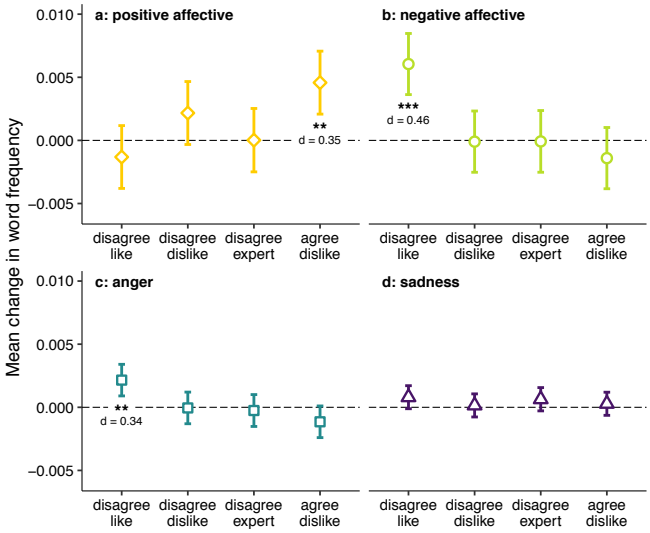

Figure 3: Mean change (post- minus pre-treatment) in frequency of affective LIWC word classes in texts written by respondents to explain their opinions towards topics. Positive values denote increases in word frequency. Asterisks denote significance levels of FDR-corrected p-values ($^{*}p_{\text{corr}} < 0.05$, $^{**}p_{\text{corr}} < 0.01$, $^{***}p_{\text{corr}} < 0.001$). Cohen's d is shown for all significant effects.

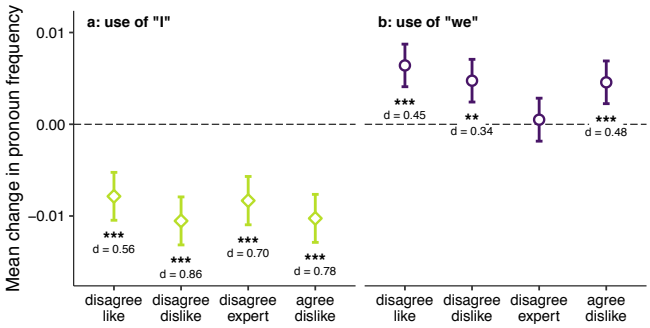

Figure 4: Mean change (post- minus pre-treatment) in frequency of first-person pronouns in texts written by respondents to explain their attitudes towards topics. Positive values denote increases in pronoun usage. Asterisks denote significance levels of FDR-corrected p-values ($^{*}p_{\text{corr}} < 0.05$, $^{**}p_{\text{corr}} < 0.01$, $^{***}p_{\text{corr}} < 0.001$). Cohen's d shown for all significant effects.

In contrast, the results showed a slight overall increase in the change of use of words that indicate sadness, as shown in Fig. 3d, but this change was extremely small (Cohen's $d = 0.06$) and not significant ($F(3, 1480) = 0.422$, $p = 0.737$).

Exploratory analysis: use of pronouns. In our analysis of LIWC words, we also explored the change in respondents' use of pronouns

in their opinion statements before and after treatment. Specifically, we observed a significant decrease in the use of the first-person singular pronoun "I" for all spokesperson conditions, while we simultaneously found an increase in the first-person plural pronoun "we" for all spokesperson conditions except disagree/expert, for which we observed almost no change (Fig. 4). While there could be mundane explanations for this effect, such as the fact that the respondents were tasked to engage with another person's opinion, it could also be indicative of an increased sense of community as a result of the task and the topics. Furthermore, given the sensitive and polarizing nature of the four topics, this shift is also consistent with the change in language as a reaction to a situation that induces trauma or shock, which have been shown to result in a decrease in use of pronouns that reflect the self (i.e., I-words) and an increase in use of pronouns that convey a sense of community (i.e., we-words) [45].

6 DISCUSSION

In the following, we discuss the implications of our results for the representation of opinions on polarizing topics in the online world.

6.1 The Impossibility of Opinion Change

In contrast to our initial goal for the RCT, in which we aimed to soften extreme opinions by employing liked celebrities, we found that changing the opinions that respondents held by confronting them with an opposite opinion was unsuccessful. If a celebrity disagreed with the respondent's prior opinion towards the topic, the attitude towards the celebrity (i.e., whether they liked or disliked the celebrity) had no significant impact on the shift in opinion. In the case of the expert spokesperson, however, we found that disagreement led to a fortification effect in which respondents became further entrenched in their own prior opinion. While consistent with previous work showing that individuals tend to distance themselves from opposing opinions and take a more extreme position [36], this effect in our study is specific to the disagree/expert condition, which puts into question the perceived influence of experts and their ability to mediate in such situations. One potential explanation for the observed rejection of a counter-attitudinal expert opinion in particular, as opposed to the disagree/like and disagree/dislike conditions, we speculate, is due to the fact that only the expert's opinion has the potential to invalidate the respondent's opinion on a scientific basis. Thus, our findings provide experimental evidence of a backfire effect [43], but only under this specific condition, and extend previous research showing that there are tendencies in partisan environments to reject the correct interpretation of scientific results when that interpretation conflicts with previously held opinions [57].

When considering the effects of being exposed to the opinion of a disliked celebrity, disagreement led to no change in the respondent's opinion. On the other hand, however, agreement by a similarly disliked celebrity led to entrenchment towards the respondent's previously held own opinion, in what seems to be a justification effect (i.e., along the line of thinking "If even this person whom I cannot stand agrees with me, how could I possibly be wrong?"). This entrenchment is consistent with studies showing that people generally surround themselves with those with whom they agree, in part because receiving validation for one's view of the world is reinforcing [26]. Overall, none of the spokesperson conditions were effective in mellowing extreme opinions, while two of the spokesperson conditions—the disagree/expert and agree/dislike conditions—led to entrenchment.

6.2 The Implications of Empathy Change

Overall, independent of the spokesperson, we observed a significant decrease in empathy when respondents considered opinions on climate change and immigration. However, when we split the responses by the spokesperson condition, we found that this effect is likely a reaction to disagreement. Specifically, while we found that there was no change in empathy towards an agreeing spokesperson, there was a significant drop in empathy towards disagreeing spokespersons, independent of the respondent's attitude towards them (liked, disliked, or an expert who was not previously known to the respondent). Not only does this finding mirror the results of the changes in opinion, it also paints a clear image of punitive behavior, in which respondents displayed less empathy towards spokespersons who disagreed with their own opinion. Remarkably, the agreeing position of the disliked celebrity seems to confer immunity from potential punitive behavior. However, the implications for the exploitation of Internet tribalism are unfortunately still obvious when we consider that none of the spokespersons ever truly uttered any of the quotations that we used in this study. Thus, it seems feasible to decrease someone's attitude towards a third party simply by falsely portraying them to hold a dissenting view.

6.3 The Implications of Language Change

While we only found the changes in affective language to be significant for two spokesperson conditions (an increase in positive affective language for agree/dislike and an increase in negative affective language for disagree/like), these findings again mirror the change in empathy towards the spokespersons. Thus, they validate the assumption that agreement by a disliked spokesperson will create a more positive inclination, while disagreement by a liked spokesperson creates a negative inclination. The latter might lead to a decrease in attraction towards the celebrity, which in turn might limit the opportunity for future positive influence.

In particular, the increase in anger is substantial, which suggests that simply encountering disagreement on a polarizing opinion, including on the Web, can increase the likelihood of an angry response and could serve as an explanation for the ease with which discussions escalate. This is consistent with previous research showing that verbal aggressiveness is a negative predictor of tolerance for disagreement [34].

6.4 The View from Above and Beyond

In considering our collective results for opinion change, empathy change, and language change, we find strong and converging evidence that disagreement has no positive effect: (*i*) respondents' opinions did not change towards the viewpoint of a disagreeing spokesperson, even when it came from a liked celebrity; (*ii*) respondents' empathy decreased towards disagreeing spokespersons, indicating a punitive effect; and (*iii*) in the case of a liked celebrity, a disagreeing opinion even increased the observed level of negative

emotions in general, and anger in particular, that were expressed in the respondents' language. In contrast, the effects of agreement by a disliked celebrity did not result in a change in empathy but resulted in a hardening of the previously held opinion and more positive language use, which may indicate the possibility of generating an increase in empathy through sustained interaction.

Particularly troubling is the backfire effect of the position held by the expert, which raises to question the role that experts can play in counteracting misinformation and unscientific beliefs in the online world, where they are frequently shared (consider, for example, that climate change and vaccination—two of the topics in this study—are prototypical examples of such misguided beliefs [29, 31]). In addition, the decrease in empathy for the expert, which was substantial and comparable to the disagree/like celebrity condition, may indicate a joint sense of disappointment in both experts and liked celebrities when they offer a dissenting point of view.

With regard to the small observed effect sizes, we emphasize that the study contained only minimal interaction with the quotation during treatment for reasons of feasibility, lasting only a few minutes (or less) per topic. However, even this minimal interaction still yielded a significant effect that cannot be discounted—and in real-world scenarios, a small effect may be all that is needed. Consider, for example, the small margins by which democratic elections tend to be decided and the enormous consequences that even a small effect might have on the outcome. Given our central finding that it is much easier to foment dissent, achieve entrenchment, and decrease empathy, it is clear that this warrants further investigation.

6.5 Limitations

Given the complex setup of the RCT and the design decisions that were necessary to make this study feasible, it is clear that it has several limitations, which we discuss in the following.

Spokesperson conditions. The original aim of the RCT was to investigate the softening of opinions by employing spokespersons that differed in terms of likeability (like, dislike, and unknown expert) and (dis)agreement, resulting in six possible combinations. We opted to only use four out of the six combinations since the inclusion of two additional spokesperson conditions would have entailed the inclusion of two additional topics and thereby increased the size of the survey (and the cost) by about half. Consequently, the two spokesperson conditions agree/like and agree/expert are missing in our results. While this does not invalidate our findings, these conditions would be useful in obtaining the full picture, especially with regard to the changes in empathy towards liked spokespersons. In particular the justification effect (i.e., the reinforcing of one's already-held opinion due to an agreeing, disliked spokesperson) would benefit from further consideration by directly comparing the effect between liked and disliked spokespersons that agree with the respondent's opinion.

Spokesperson selection. For reasons of feasibility, we also did not design the study to investigate the effect of gender or ethnicity on opinion or empathy change. We selected celebrity spokespersons in a data-driven manner (see Appendix A.1), which led to an under-representation of women among the celebrity spokespersons (20% female celebrities, even though over 50% of crowd workers were female). Due to the lack of ethnically diverse images for the empathy-for-pain test, we also did not include spokespersons with different skin colors. Future studies could focus on either variable. A further variable that could be considered is a celebrity's convincingness throughout their career [11]. Finally, since we wanted to ensure comparability of the findings for the expert condition across topics (which we expected to perform well in softening extreme opinions), we used the same fictitious expert for all topics. In contrast to our results, a known medical expert spokesperson was found to increase the engagement of social media users for the topic of physical distancing during the early stages of the COVID-19 pandemic [2]. Thus, an investigation into the performance of different experts (including gender and ethnicity) and of real, known experts would be of interest to determine the factors that might make experts more convincing. Such future research should also consider the potential effects that could come into play in crisis versus non-crisis situations in determining the performance of experts, and spokespersons more generally.

Topic selection. Due to the expectedly small effect sizes and the substantial financial overhead of running a survey with even this number of respondents, we intentionally limited the selection of topics to those that are highly polarizing in the current political climate. In subsequent, narrower studies that allow for a larger number of respondents, the potential for opinion change in less extreme settings with a lower risk of tribalism-related effects could be considered.

Long-term change. We focused on short-term changes in opinion and empathy for this prototypical experimental study. Thus, it is unclear to which degree the effects endure over time or if lasting change requires additional boosting. While the observed short-term effects already stand to have a substantial impact in the context of societal decision with small decision margins, a longitudinal study is required to assess the long-term effects of spokesperson selection on opinion and empathy change.

7 CONCLUSION AND OUTLOOK

With the goal of shedding more light on the phenomenon of polarization, we investigated the potential influence of celebrity and expert opinions on softening the extreme views of respondents on topics known to ignite polarized debates in society. The results of our RCT suggest that exposing respondents to opposing views of liked celebrities is ineffective in mellowing their extreme opinions, and—contrary to expectations—even leads to the expression of punitive behavior and the arousal of anger and negative feelings. The present research also speaks to a broader concern about science denial as is evident from the backfire effect and respondents' punitive response induced by exposure to the disagreeing opinion of the scientific expert. Thus, our findings suggest that spokespersons may not only fail to have the intended effect of softening opinions when their views are too different from the respondents, but that they may actually backfire by moving people in the opposite direction.

Intriguingly, we also observed opinion hardening when respondents were exposed to an agreeing opinion of a celebrity whom they disliked, which was mirrored in the use of positive affective language. A practical implication of this result is that it could still

serve as a starting point in devising interventions for softening polarized opinions on the Web. In our study, we examined change as a consequence of dyadic interactions between a spokesperson and a respondent. It would be intriguing to investigate whether the influence of a spokesperson's view on the respondent's opinion might be facilitated or mediated in the presence of a confirmatory opinion from a third party, with whom the respondent knows the spokesperson to be in tension. Such *triadic interaction* might help to mitigate disagreements and control polarized debates. Extending our knowledge of this complex opinion-change process is timely in the face of increasing hostility in interactions on the Web, and has the potential to identify effective strategies in making the Web a more civil place for dialogue.

Acknowledgments. This work was partly supported by Collaborative Research on Science and Society (CROSS), Swiss Data Science Center, Swiss National Science Foundation (grant 200021_185043), Microsoft Swiss Joint Research Center. We also gratefully acknowledge generous gifts from Facebook and Google.

REFERENCES

[1] Alan I Abramowitz and Kyle L Saunders. 2008. Is Polarization a Myth? *The Journal of Politics* 70, 2 (2008), 542–555.
[2] Ahmad Abu-Akel, Andreas Spitz, and Robert West. 2021. The Effect of Spokesperson Attribution on Public Health Message Sharing During the COVID-19 Pandemic. *PLOS ONE* 16, 2 (2021), e0245100.
[3] Hunt Allcott and Matthew Gentzkow. 2017. Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives* 31, 2 (2017), 211–236.
[4] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, M B Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to Opposing Views on Social Media can Increase Political Polarization. *Proceedings of the National Academy of Sciences* 115, 37 (2018), 9216–9221.
[5] Delia Baldassarri and Andrew Gelman. 2008. Partisans Without Constraint: Political Polarization and Trends in American Public Opinion. *Amer. J. Sociology* 114, 2 (2008), 408–446.
[6] Michael D. Basil. 1996. Identification as a Mediator of Celebrity Effects. *Journal of Broadcasting & Electronic Media* 40, 4 (1996), 478–495.
[7] Christina S Beck, Stellina M Aubuchon, Timothy P McKenna, Stephanie Ruhl, and Nathaniel Simmons. 2014. Blurring Personal Health and Public Priorities: An Analysis of Celebrity Health Narratives in the Public Sphere. *Health Communication* 29, 3 (2014), 244–256.
[8] Kevin D Bird and Dusan Hadzi-Pavlovic. 2014. Controlling the Maximum Familywise Type I Error Rate in Analyses of Multivariate Experiments. *Psychological Methods* 19, 2 (2014), 265–280.
[9] Emile Bruneau, Nour Kteily, and Emily Falk. 2018. Interventions Highlighting Hypocrisy Reduce Collective Blame of Muslims for Individual Acts of Violence and Assuage Anti-Muslim Hostility. *Personality and Social Psychology Bulletin* 44, 3 (2018), 430–448.
[10] Emile G Bruneau, Nour S Kteily, and Ana Urbiola. 2020. A Collective Blame Hypocrisy Intervention Enduringly Reduces Hostility Towards Muslims. *Nature Human Behaviour* 4, 1 (2020), 45–54.
[11] François A. Carrillat and Jasmina Ilicic. 2019. The Celebrity Capital Life Cycle: A Framework for Future Research Directions on Celebrity Endorsement. *Journal of Advertising* 48 (2019), 61–71.
[12] Rhia Catapano, Zakary L Tormala, and Derek D Rucker. 2019. Perspective Taking and Self-Persuasion: Why "Putting Yourself in Their Shoes" Reduces Openness to Attitude Change. *Psychological Science* 30, 3 (2019), 424–435.
[13] Rhonda Christensen and Gerald Knezek. 2015. The Climate Change Attitude Survey: Measuring Middle School Student Beliefs and Intentions to Enact Positive Environmental Change. *International Journal of Environmental and Science Education* 10, 5 (2015), 773–788.
[14] Robert B Cialdini. 2007. *Influence: The Psychology of Persuasion.* Collins New York.
[15] Smiljana J Cvjetkovic, Vida Lj Jeremic, and Danijela V Tiosavljevic. 2017. Knowledge and Attitudes Toward Vaccination: A Survey of Serbian Students. *Journal of Infection and Public Health* 10, 5 (2017), 649–656.
[16] Dominic DiFranzo and Marie Joan Kristine Gloria. 2017. Filter Bubbles and Fake News. *XRDS* 23, 3 (2017), 32–35.
[17] B Zafer Erdogan, Michael J Baker, and Stephen Tagg. 2001. Selecting Celebrity Endorsers: The Practitioner's Perspective. *Journal of Advertising Research* 41, 3 (2001), 39–48.
[18] Lawrence Ezrow, Margit Tavits, and Jonathan Homola. 2014. Voter Polarization, Strength of Partisanship, and Support for Extremist Parties. *Comparative Political Studies* 47, 11 (2014), 1558–1583.
[19] James Fishkin, Alice Siu, Larry Diamond, and Norman Bradburn. 2020. Is Deliberation an Antidote to Extreme Partisan Polarization? Reflections on America in One Room. *APSA Preprints* (2020).
[20] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. 2018. Reducing Controversy by Connecting Opposing Views. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 5249–5253.
[21] Gordon Gauchat. 2012. Politicization of Science in the Public Sphere: A Study of Public Trust in the United States, 1974 to 2010. *American Sociological Review* 77, 2 (2012), 167–187.
[22] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A Very Brief Measure of the Big Five Personality Domains. *Journal of Research in Personality* 29, 6 (2003), 504–528.
[23] Boaz Hameiri, Daniel Bar-Tal, and Eran Halperin. 2019. Paradoxical Thinking Interventions: A Paradigm for Societal Change. *Social Issues and Policy Review* 13, 1 (2019), 36–62.
[24] Boaz Hameiri, Roni Porat, Daniel Bar-Tal, Atara Bieler, and Eran Halperin. 2014. Paradoxical Thinking as a New Avenue of Intervention to Promote Peace. *Proceedings of the National Academy of Sciences* 111, 30 (2014), 10996–11001.
[25] Joshua A Hemmerich, Kellie Van Voorhis, and Jennifer Wiley. 2016. Anomalous Evidence, Confidence Change, and Theory Change. *Cognitive Science* 40, 6 (2016), 1534–1560.
[26] Clayton J Hilmert, James A Kulik, and Nicholas J S Christenfeld. 2006. Positive and Negative Opinion Modeling: The Influence of Another's Similarity and Dissimilarity. *Journal of Personality and Social Psychology* 90, 3 (2006), 440–452.
[27] David J. Jackson. 2018. The Effects of Celebrity Endorsements of Ideas and Presidential Candidates. *Journal of Political Marketing* 17, 4 (2018), 301–321.
[28] Philip L Jackson, Andrew N Meltzoff, and Jean Decety. 2005. How Do We Perceive the Pain of Others? A Window Into the Neural Processes Involved in Empathy. *NeuroImage* 24, 3 (2005), 771–779.
[29] Daniel Jolley and Karen M Douglas. 2014. The Effects of Anti-Vaccine Conspiracy Theories on Vaccination Intentions. *PLOS ONE* 9, 2 (2014), 1–9.
[30] Daniel Jolley, Rose Meleady, and Karen M Douglas. 2020. Exposure to Intergroup Conspiracy Theories Promotes Prejudice Which Spreads Across Groups. *British Journal of Psychology* 111, 1 (2020), 17–35.
[31] Stephan Lewandowsky, Gilles E Gignac, and Klaus Oberauer. 2015. The Robust Relationship Between Conspiracism and Denial of (Climate) Science. *Psychological Science* 26, 5 (2015), 667–670.
[32] Stephan Lewandowsky, Gilles E Gignac, and Samuel Vaughan. 2013. The Pivotal Role of Perceived Scientific Consensus in Acceptance of Science. *Nature Climate Change* 3, 4 (2013), 399–404.
[33] Qingzi Vera Liao and Wai-Tat Fu. 2014. Can You Hear Me Now?: Mitigating the Echo Chamber Effect by Source Position Indicators. In *Computer Supported Cooperative Work (CSCW)*. 184–196.
[34] Darren L Linvill, Joseph P Mazer, and Brandon C Boatwright. 2016. Need for Cognition as a Mediating Variable Between Aggressive Communication Traits and Tolerance for Disagreement. *Communication Research Reports* 33, 4 (2016), 363–369.
[35] Lilliana Mason. 2013. The Rise of Uncivil Agreement: Issue Versus Behavioral Polarization in the American Electorate. *American Behavioral Scientist* 57, 1 (2013), 140–159.
[36] Winter A Mason, Frederica R Conrey, and Eliot R Smith. 2007. Situating Social Influence Processes: Dynamic, Multidirectional Flows of Influence Within Social Networks. *Personality and Social Psychology Review* 11, 3 (2007), 279–300.
[37] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.
[38] Jaroslav Mihálik and Matúš Jankola. 2016. European Migration Crisis: Positions, Polarization and Conflict Management of Slovak Political Parties. *Baltic Journal of Law & Politics* 9, 1 (2016), 1–25.
[39] Shekhar Misra and Sharon E Beatty. 1990. Celebrity Spokesperson and Brand Congruence: An Assessment of Recall and Affect. *Journal of Business Research* 21, 2 (1990), 159–173.
[40] Aaron J Moss and Leib Litman. 2018. After the Bot Scare: Understanding What's Been Happening With Data Collection on MTurk and How to Stop It. *Cloud Research* (2018). https://www.cloudresearch.com/resources/blog/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it
[41] Sean A. Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *International Conference on Human Factors in Computing Systems (CHI)*. 1457–1466.
[42] Raymond S Nickerson. 1998. Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology* 2, 2 (1998), 175–220.
[43] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[44] Dario Pavllo, Tiziano Piccardi, and Robert West. 2018. Quootstrap: Scalable Unsupervised Extraction of Quotation-Speaker Pairs from Large News Corpora via Bootstrapping. In *International AAAI Conference on Web and Social Media (ICWSM)*.
[45] James W Pennebaker. 2013. *The Secret Life of Pronouns: What Our Words Say About Us.* Bloomsbury Publishing.
[46] 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 (2019), 39–50.
[47] Daniel K Pryce. 2018. U.S. Citizens' Current Attitudes Toward Immigrants and Immigration: A Study From the General Social Survey. *Social Science Quarterly* 99, 4 (2018), 1467–1483.
[48] Simone G Shamay-Tsoory, Ahmad Abu-Akel, Sharon Palgi, Ramzi Sulieman, Meytal Fischer-Shofty, Yechiel Levkovitz, and Jean Decety. 2013. Giving Peace a Chance: Oxytocin Increases Empathy to Pain in the Context of the Israeli–Palestinian Conflict. *Psychoneuroendocrinology* 38, 12 (2013), 3139–3144.
[49] Bruno Castanho Silva. 2018. Populist Radical Right Parties and Mass Polarization in the Netherlands. *European Political Science Review* 10, 2 (2018), 219–244.
[50] Károly Takács, Andreas Flache, and Michael Mäs. 2016. Discrepancy and Disliking Do Not Induce Negative Opinion Shifts. *PLOS ONE* 11, 6 (2016), 1–21.
[51] Yla R Tausczik and James W Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.
[52] Michael G Taylor and George I Whitehead. 2014. The measurement of attitudes toward abortion. *Modern Psychological Studies* 20, 1 (2014), 79–86.
[53] John C Turner. 1991. *Social Influence.* Thomson Brooks/Cole Publishing.
[54] Petter Törnberg. 2018. Echo Chambers and Viral Misinformation: Modeling Fake News as Complex Contagion. *PLOS ONE* 13, 9 (2018), 1–21.
[55] Sander van der Linden, Anthony Leiserowitz, and Edward Maibach. 2018. Scientific Agreement can Neutralize Politicization of Facts. *Nature Human Behaviour* 2, 1 (2018), 2–3.
[56] Jan-Willem van Prooijen, André P M Krouwel, and Thomas V Pollet. 2015. Political Extremism Predicts Belief in Conspiracy Theories. *Social Psychological and Personality Science* 6, 5 (2015), 570–578.
[57] Anthony N Washburn and Linda J Skitka. 2018. Science Denial Across the Political Divide: Liberals and Conservatives Are Similarly Motivated to Deny Attitude-Inconsistent Science. *Social Psychological and Personality Science* 9, 8 (2018), 972–980.
[58] Thomas Wood and Ethan Porter. 2019. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior* 41, 1 (2019), 135–163.

A STUDY DETAILS

A.1 Spokesperson Selection

Celebrity candidate elicitation. Since the proper definition of what constitutes a celebrity is vague, we determined the pool of celebrity spokespersons in a data-driven manner by collecting from crowd workers on Amazon Mechanical Turk the names of celebrities, from various domains, whom they liked and disliked in a preparatory survey. We elicited celebrity names from the areas of (i) science and education, (ii) sports, (iii) politics, (iv) movies and TV, and (v) music. The survey was given to 103 crowd workers that specified their residence to be in the US. Respondents were also asked to input their two-letter state code. All responses were manually checked for quality to avoid the inclusion of responses from non-native crowd workers connecting via VPN and unfamiliar with US culture. Seven response sets were discarded due to incomplete data, low quality, or invalid state codes. Since responses were elicited as free text and thus extremely noisy, the results were cleaned semi-automatically by applying multiple iterations of string matching with a human annotator in the loop.

Celebrity popularity elicitation. Using the above pool of names, we next determined the popularity and likeability of celebrities to select spokespersons for the RCT. We discarded celebrity candidates from the pool of science and education due to the overall low number of responses in this area, indicating a low level of popularity. Celebrities in the remaining four pools were ranked by mention frequency, and the twelve most frequently named celebrities were selected for each pool (48 celebrities in total). We used a second preparatory survey to determine (i) the general level of likeability and (ii) the public knowledge of these celebrities' opinions towards the four topics, to ensure that their opinions were sufficiently unknown to attribute quotations from both ends of the polarity spectrum. Respondents were asked to state whether they liked, disliked, were ambivalent towards a celebrity, or did not know them. For each celebrity and topic, respondents were also asked to state, to the best of their knowledge, on which end of the polarity spectrum the opinion of the celebrity was located. We collected responses from 150 crowd workers per celebrity. The quality of the responses was satisfactory, so no data had to be discarded.

Celebrity spokesperson selection. Based on the collected popularity and opinion data, we then computed the unknown-opinion ratio for each celebrity–topic combination as the fraction of respondents who did not know this celebrity's opinion on the given topic. After filtering celebrities based on a lower threshold θ on the unknown-opinion ratio, we ranked celebrities by their aggregated ratio over all four topics. Using this ranking, we then iteratively removed celebrities from the pool (starting with the celebrity with the lowest score) until further removal would have resulted in instances of respondents who could not be assigned a celebrity with (to them) unknown opinion. This point was reached at a pool size of 25 celebrities for three different threshold values $\theta = 0.5, 0.6, 0.7$ (with higher threshold values leading to no celebrity candidates being viable and lower threshold values causing all candidates to be viable). We selected these 25 celebrities for inclusion in the randomized controlled trial (for a list of the celebrity spokesperson candidates, see Appendix D). During this selection process, we found that all celebrities from the area of politics were discarded due to their opinions being well known to a majority of respondents.

Crowd workers who participated in any preparatory survey were excluded from participating in the RCT in order to avoid bias.

Spokesperson portraits. We collected portraits of the 25 celebrity spokespersons from the Web. To ensure comparability between the reactions of respondents who were assigned different spokespersons in the RCT, we selected all portraits according to the following criteria: (i) portraits were selected to have a neutral background, (ii) the spokesperson was facing the camera, and (iii) the facial expression of the spokesperson was neutral.

Expert spokesperson. To ensure that the expert spokesperson was not known to respondents of the RCT and could be assigned to all four topics, we created a fictitious expert. We used the image of Prof. Michael Rosbash, who was awarded the Nobel Prize in Medicine in 2017, to create a likeness for this fictitious expert. To reduce the risk of recognition by the crowd workers, we assigned the expert the fake name "Dr. Michael Barnes", which did not result in noteworthy Google results in combination with any of the four topics at the time of the study.

A.2 Quotation Selection

For each of the two polarities per topic, one quotation was selected to represent this viewpoint (for a total of 8 quotations). Quotations were manually chosen to represent opposite ends of the spectrum on this topic (the full text of these quotations can be found in Appendix C). For the extraction of a large pool of contemporary quotations from which we could choose, we applied the quotation extraction and attribution tool Quootstrap [44] to a corpus of 129 million news articles provided by the online content aggregation service Spinn3r.com. We pre-filtered quotations by keyword searches for each topic, before candidates were manually screened to identify suitable quotations.

A.3 Study Deployment

Deployment time frame. The collection of celebrity names was conducted from 27 August to 3 September 2019. Celebrity likeability and perceived outspokenness about each of the four topics was determined on 22 and 23 September 2019. The screening survey was conducted between 25 and 30 October 2019. Data for the RCT was collected between 1 and 27 November 2019.

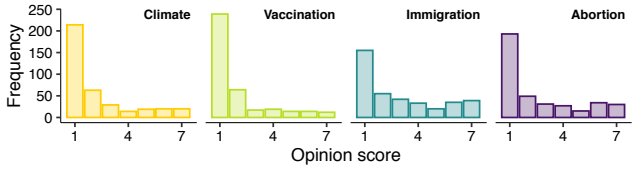

Figure 5: Frequency of respondents' pre-treatment opinions on the four topics on a 7-point Likert scale, ranging from a strong liberal view (1) to a strong conservative view (7) for abortion, climate change, and immigration. For vaccination, 1 corresponds to a pro-vaccination view.

Respondent selection. To identify viable respondents for the RCT, we used the screening survey as described in Section 3.2. Since India is the second largest source of crowd workers outside of the US and Indian crowd workers have used VPN connections to pose as US workers in the past [40], we also included three Indian celebrities in the screening survey as a honeypot to identify and exclude crowd workers with residence outside of the US (the Indian celebrities were Virat Kohli, Salman Khan, Sonam Kapoor). However, no workers had to be excluded since no worker was familiar with more than one of these celebrities.

Based on their responses to the screening survey, respondents were identified as eligible for the RCT if they liked at least one and disliked at least two of the 25 spokesperson candidates (to match the trial conditions), provided a valid US state code, were 18–68 years old, and stated in the screening survey that they would like to participate in a follow-up survey. Due to the writing-intensive treatment step in the RCT, workers were also excluded if their response to a question about daily news consumption was less than a sentence long or entirely empty (note that we only considered length, not content, as a criterion, and sentences such as "I do not consume any news" were considered valid).

Sample size. We determined the required sample size with a power analysis. Assuming a small effect size $f = 0.1$, $\alpha = 0.05$, and a power of 0.80 for one group and four conditions, we expected to need responses from 366 respondents. Based on the respondent selection process, 499 crowd workers qualified and were offered a customized survey for the randomized controlled trial. Out of these, 379 workers completed the main survey.

A.4 Ethical Considerations

Approval for this research project was obtained from the EPFL Human Research Ethics Committee. Before they took the survey, all respondents were informed that their responses would be used as part of a research project. After completing the survey, respondents received an explanation of the true aim of our research project, were debriefed on the misattribution of quotations, and shown the original source for each quotation (see Appendix C). All respondents were offered the option of dropping out of the study after completing the survey, but no respondent made use of this.

B POLITICAL LEANING OF PARTICIPANTS

Based on the opinions elicited during the pre-treatment phase, we found that the majority of our respondents—crowdworkers on Amazon Mechanical Turk—held liberal views on all topics. Specifically, 72% of respondents had a liberal or a leaning-liberal opinion on abortion, 80% on climate change, 66% on immigration, and 84% on vaccination (for the full distribution, as well as a note on vaccination, see Fig. 5). Our findings can therefore be interpreted as an observation of a backlash in a pool of respondents who predominantly hold liberal opinions, whom one would expect to be more tolerant ("liberal" in the literal sense) in their reaction to an opposing opinion. While we do not see a reason why this skewness would make the results less relevant, future research could leverage a stratified sample to ensure sufficient power to separately identify effects on both sides of the political spectrum.

C QUOTATIONS

To be able to use quotations at both ends of the polarity spectrum, we falsely attributed quotations to the spokespersons. The following is a list of all used quotations and their actual sources.

Climate Change

(+) "Climate change is a natural phenomenon. Carbon dioxide is not a pollutant. On the contrary, it makes crops and forests grow faster. Economic analysis has demonstrated that more CO_2 and a warmer climate will raise the gross national product and therefore the average income." – Fred Singer

(–) "The impact of climate change is clear, from rising sea levels to more powerful and frequent extreme weather events that put our families and businesses at risk. We have an obligation to address the root cause of these changes and that means limiting carbon pollution from our power plants." – Jack Markell

Vaccination

(+) "People think vaccines cause autism, among other things. This is not true. This is a complete myth and parents need to know that vaccines will protect their children against diseases like measles, but also that they have to vaccinate children to protect the community as well." – Leana Wen

(–) "The medical authorities keep lying. Vaccination has been a disaster on the immune system. It actually causes a lot of illnesses. We are actually changing our genetic code through vaccination... 100 years from now we will know that the biggest crime against humanity was vaccines." – Guylaine Lanctot

Immigration

(+) "Immigrants expand the U.S. economy's productive capacity, stimulate investment, and promote specialization that in the long run boosts productivity. There is no evidence that these effects take place at the expense of jobs for workers that are born here." – Giovanni Peri

(–) "It is time to stand strong for the American people. After years of mass immigration, falling wages, and surging joblessness, isn't it time we focused on the needs of the people that were born here? Isn't it time we got our own people back to work?" – Jeff Sessions

Abortion

(+) "Amid a nationwide attack on those who seek and provide abortions, the U.S. Supreme Court sided with the Constitution, and for that we should all be thankful. Abortion is a legal, constitutionally protected right that should be available to all women." – Daylin Leach

(–) "I'm pro-life and I don't apologize for it. I believe it is morally abhorrent to end an innocent human life, and I also believe it is morally wrong to use the tax dollars of millions of pro-life Americans to fund research that involves the destruction of human embryos." – Mike Pence

D CELEBRITY SPOKESPERSONS

In the RCT, we used the following 25 celebrities from film and TV (F), music (M), and sports (S) as spokesperson candidates:

Aaron Rodgers (S), Adam Sandler (F), Adele (M), Anne Hathaway (F), Ariana Grande (M), Ben Roethlisberger (S), Brad Pitt (F), Derek Jeter (S), Elton John (M), Eminem (M), Jerry Jones (S), Julia Roberts (F), Justin Bieber (M), Kevin Spacey (F), Kim Kardashian (F), Michael Phelps (S), Nicolas Cage (F), Peyton Manning (S), Quentin Tarantino (F), Robert De Niro (F), Stephen Curry (S), Tim Tebow (S), Tom Brady (S), Tom Cruise (F), Tom Hanks (F).