Information Processing and Management 62 (2025) 104244

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

# Enhancing media literacy: The effectiveness of (Human) annotations and bias visualizations on bias detection

Timo Spinde [a,*], Fei Wu [b], Wolfgang Gaissmaier [c,d], Gianluca Demartini [e], Isao Echizen [f], Helge Giese [c,g]

[a] Department of Computer Science, University of Göttingen, Göttingen, Germany
[b] Institute for Natural Language Processing, University of Stuttgart, Stuttgart, Germany
[c] Department of Psychology, University of Konstanz, Konstanz, Germany
[d] Cluster for the Advanced Study of Collective Behavior, University of Konstanz, Konstanz, Germany
[e] School of Electrical Engineering and Computer Science, The University of Queensland, Brisbane, Australia
[f] National Institute of Informatics, Tokyo, Japan
[g] Heisenberg Chair for Medical Risk Literacy and Evidence-Based Decisions, Charité – Universitätsmedizin Berlin, Berlin, Germany

ARTICLE INFO

*Keywords:*
News literacy
Media bias
Language processing
Text perception

ABSTRACT

Marking biased texts effectively increases media bias awareness, but its sustainability across new topics and unmarked news remains unclear, and the role of AI-generated bias labels is untested. This study examines how news consumers learn to perceive media bias from human- and AI-generated labels and identify biased language through highlighting, neutral rephrasing, and political orientation cues. We conducted two experiments with a teaching phase exposing them to various bias-labeling conditions and a testing phase evaluating their ability to classify biased sentences and detect biased text in unlabeled news on new topics.

We find that, compared to the control group, both human- and AI-generated sentential bias labels significantly improve bias classification ($p < .001$), though human labels are more effective ($d = 0.42$ vs. $d = 0.23$). Additionally, among all teaching interventions, participants best detect biased sentences when taught with biased sentence or phrase labels ($p < .001$), while politicized phrase labels reduce accuracy. The effectiveness of different media literacy interventions remains independent of political ideology, but conservative participants are generally less accurate ($p = .011$), suggesting an interaction between political inclinations and bias detection.

Our research provides a novel experimental framework into assessing the generalizability of media bias awareness and offer practical implications for designing bias indicators in news-reading platforms and media literacy curricula.

## 1. Introduction

### 1.1. Background

News media play a central role in informing the public about crucial societal issues. In areas where audiences do not have direct knowledge or experience, the audiences become particularly reliant on the media that transmit the information (Happer & Philo,

* Corresponding author at: University of Göttingen, Papendiek 14 37073 Göttingen, Germany.
*E-mail address:* t.spinde@media-bias-research.org (T. Spinde).

https://doi.org/10.1016/j.ipm.2025.104244
Received 10 December 2024; Received in revised form 21 April 2025; Accepted 30 May 2025
Available online 12 June 2025

2013). News media often introduce biases by simplifying complex scientific or social issues to create a storyline that engages the audience (Henderson & Green, 2020). The term "media bias" refers to such slanted news coverage or other biased media content (Hamborg et al., 2019; Spinde, 2021) that favors a particular perspective, ideology, or result (Williams, 1975). During the news production process, bias can be introduced in various ways, such as through journalists' choices of biased words and framing of a topic, or editorial decisions (Hinterreiter et al., 2025).

Media bias significantly impacts the public's knowledge of specific issues and decision-making processes, especially when news consumers are unaware of the extent of the bias (Druckman & Parkin, 2005; Eberl et al., 2017; Spinde et al., 2024). For instance, during the pandemic of COVID-19, L.Y. found those who had greater knowledge of the pandemic were more likely to be critical of biased media content when formulating their own opinions, while those who relied only on certain news sources were more likely to be influenced unconsciously. Furthermore, biased media content can lead to polarization in society and foster the audience's distrustful attitude toward the authorities (L.Y. Dhanani & Franz, 2020). Media bias also impacts other domains such as politics (Bernhardt et al., 2008; Eberl et al., 2017; Islam, 2008), economics (McCarthy & Dolfsma, 2014), environment (Henderson & Green, 2020), and social safety issues (Sugimoto et al., 2013). These observations highlight a need for strategies that mitigate the negative effects of media bias, especially in cases of controversial topics that are societally relevant.

Furthermore, the rapid news production cycle on digital platforms has amplified the amount of biased content exposed to news consumers (Iandoli et al., 2021; Terren & Borge-Bravo, 2021). Customized news recommendation systems also contribute to echo chambers (Auxier & Vitak, 2019), where news consumers are more appealed to read news aligning with their opinions, and their biased perspectives are further reinforced (Spinde et al., 2024). As the volume of digital information continues to grow, obtaining an overview of bias across media outlets is only feasible through automated solutions (Spinde et al., 2021c; Wessel et al., 2023).

### 1.2. Related work

Previous research has explored mitigating media bias by increasing readers' awareness through visualizations that highlight biased text (Spinde et al., 2020, 2022) or present diverse viewpoints to promote balanced news consumption (Joris et al., 2024; Mattis et al., 2024; Munson et al., 2021; Paramita et al., 2024; Park et al., 2009, 2012; Rieger et al., 2024). While these studies suggest that visualizations help consumers recognize bias, they do not quantitatively assess their effectiveness in sustaining media bias awareness once visual aids are removed. Additionally, research has yet to evaluate the effectiveness of different media literacy interventions in teaching readers to actively detect bias at the word level.

Research also explored using natural language processing (NLP) and machine learning to automatically detect media bias (He et al., 2021; Hube & Fetahu, 2019; Lee et al., 2022; Lei et al., 2022; Liu et al., 2021; Pryzant et al., 2020; Spinde, 2021). While their results indicate that scalable, automated methods are possible for handling the amount of daily news content (Hamborg et al., 2019; Spinde et al., 2024), the effectiveness of visualizations based on automatically generated bias labels in increasing readers' bias awareness remains untested. Prior visualization experiments rely on manual annotations, limiting scalability (Spinde et al., 2020, 2022).

We define "media bias awareness" as a reader's ability to perceive bias in news content without explicitly identifying specific biased elements. Our study further investigates participants' ability to detect biased language in news content, assessing the extent to which they can recognize media bias. Throughout this paper, we use "teaching" to refer to media literacy interventions for human participants and "training" when discussing AI model development.

### 1.3. Research questions and contributions

Our research addresses this gap through two structured experiments, incorporating teaching and testing phases to assess the effectiveness of media literacy interventions after their removal. Study 1 examines interventions from different sources (AI- and human-generated), while Study 2 evaluates various intervention types. We investigate their role in sustaining media bias awareness and enhancing participants' ability to detect biased text in unmarked news on unfamiliar topics.

In Study 1, the teaching phase exposes participants to AI- and human-generated media literacy interventions, all presented in the same form. The testing phase assesses whether media bias awareness persists when reading plain sentences after visual aids are removed and evaluates the effectiveness of AI- and human-generated labels.

In Study 2, the teaching phase introduces interventions of different types from the same sources to examine how intervention format impacts learning. The testing phase evaluates whether participants can identify biased phrases in articles on different topics after exposure to various types of media literacy interventions.

We formulate our research questions (RQ) as follows:

- RQ1: How effective are AI-generated bias labels and Human labels in sustaining participants' media bias awareness, compared to a control group without any media literacy interventions?
- RQ2: Which (combination of) media literacy intervention is the most effective in teaching participants to detect biased language post-visualization removal?
- RQ3: Do participants' backgrounds, such as political inclinations, interact with different media literacy interventions to jointly influence their learning effects?

We address RQ1 in Study 1 and RQ2 and RQ3 in Study 2. We preregistered our hypotheses at https://osf.io/q8rcu/ and https://osf.io/95ht6/, where we also make all data and code publicly available. Through these studies, we make the following contributions:

1. We propose a novel, quantitative approach to measuring the effectiveness of media literacy interventions post-visualization removal in sustaining readers' media bias awareness and their ability to detect biased language.
2. In Study 1, we explore AI-generated bias labels as a scalable alternative for media literacy interventions. While human-generated labels yield better results, AI labels still significantly enhance media bias awareness compared to the control group, highlighting their potential in bias visualizations.
3. In Study 2, we find that participants continue to detect bias in the test phase, suggesting that learning effects persist and generalize to plain news content on unfamiliar topics.
4. Study 2 also reveals that phrasal bias highlighting is the most effective intervention in enhancing participants' ability to detect biased language.
5. Finally, we examine the interaction between various interventions and political orientation in Study 2, showing that learning effects weaken when both phrasal bias and politicized word use are highlighted. This effect was more pronounced among conservative participants, who demonstrated lower accuracy in bias detection ($p = .011$), suggesting that political inclinations influence learning outcomes. These findings align with previous research indicating that echo chambers remain difficult to overcome, even with visual aids.

## 2. Related work

Media bias manifests in various ways, including linguistic choices, editorial decisions, and broader socio-political influences (Spinde et al., 2024). This study focuses on linguistic bias, which forms the basis for the word- and sentence-level visualized labels we tested. Linguistic bias refers to word choices that reinforce stereotypes or specific perceptions of events, groups, or individuals (Beukeboom & Burgers, 2017) and can appear at the word, phrase, or sentence structure level (Hinterreiter et al., 2024).

Research on mitigating media bias has explored various interventions, yet few studies quantitatively measure their immediate impact on user awareness or assess whether this awareness persists after removing visual aids. In the following sections, we review three key areas informing our experiment design: (1) media literacy interventions, (2) bias-indicating visualizations, and (3) automated media bias detection.

### 2.1. Media literacy interventions

While structured frameworks for teaching media bias are limited, extensive research on misinformation and fake news literacy highlights the effectiveness of media literacy interventions in enhancing news credibility assessment (Goodsett, 2024; Hameleers, 2022; Lu et al., 2024; Van der Meer & Hameleers, 2021). A meta-analysis by Lu et al. (2024) found that interventions such as courses, games, videos, and graphics improve fake news credibility assessments. While their methods differ from ours, these findings suggest that teaching media literacy is a promising approach to mitigating media bias.

Two key intervention strategies are debunking, which corrects misinformation after Fance, and prebunking, which warns individuals of potential manipulation before exposure (Goodsett, 2023). Prebunking strategies, such as forewarning messages, have been shown to increase skepticism and promote critical reasoning (Piksa et al., 2024), while debunking strategies, such as providing alternative narratives, improve misinformation detection. These insights informed our study design, where we incorporated a short introduction to the concept of media bias in Study 1 and neutral rephrasing of biased sentences to enhance participants' media bias awareness in the teaching phase in Study 2.

### 2.2. Bias-Indicating visualizations

One strategy for addressing media bias focuses on exposing readers to diverse perspectives (Munson et al., 2021; Park et al., 2009, 2012), such as aggregating multiple viewpoints or enhancing news diversity in recommender systems and search results (Joris et al., 2024; Mattis et al., 2024; Paramita et al., 2024; Rieger et al., 2024). While these methods aim to balance content exposure, they do not directly address bias awareness and detection at the textual level. As our study focuses on bias-indicating visualizations within individual articles, we do not elaborate on these approaches. Instead, our research examines bias-indicating visualizations that highlight biased language at the word and sentence levels.

Despite the growing interest in media bias detection, there is limited research on the effectiveness of bias-indicating visualizations in shaping readers' awareness and detection of bias. Spinde et al. (2020) examined how highlighting different content types—key facts, framing effects, and biased language—affects human bias perception of news articles, finding that highlighting biased language was the most effective. In a follow-up study, Spinde et al. (2022) evaluated additional bias indicators and found that biased language annotations with explanations had the strongest impact, while forewarnings were also effective. In contrast, political classification had no effect. However, both studies measured bias awareness through attitude shifts (Spinde et al., 2020) or self-reported perceptions of bias (Spinde et al., 2022), which limits the objectivity and quantifiability of their findings.

Building on these insights, our experiments incorporate various types of biased language highlighting at the phrasal and sentence levels and neutral sentence rephrasing to evaluate the most effective media literacy intervention. While such techniques influence bias perception, their impact on bias awareness and detection over time remains unclear. To address this gap, our study examines whether learning effects persist after visualization removal.

### 2.3. Automatic media bias detection

Given the vast amount of online news content, relying solely on manual annotations is impractical. To address this, researchers have suggested automated labeling as a scalable alternative (Wessel et al., 2023). Natural language processing (NLP) techniques can be used to automate bias detection, enabling large-scale systems for bias analysis and mitigation (Hinterreiter et al., 2024; Wessel et al., 2023; Spinde et al., 2021). Although these models have shown promising results, their training relies on manual annotations by experts or crowdsourcers. For instance, T. achieved classifier performance with a macro F1-score of 0.804 using the Bias Annotation By Experts (BABE) dataset. Despite the dataset's high quality and strong inter-rater agreement, its limited coverage of topics and timeframes restricts its scalability.

However, it remains unclear whether automated labels are as effective as human annotations in training classifiers to detect media bias and in helping readers become more aware of biased content. Study 1 aims to bridge this gap by comparing the effectiveness of AI labels versus human-made labels in improving media bias awareness among readers.

## 3. Study 1: effectiveness of human and AI-generated labels in media bias awareness

### 3.1. Participants

We recruited 512 participants on Prolific taking part in a 12-minute survey with a reimbursement of 1.50£. We used Prolific's feature of assigning a representative sample of sex, age, and ethnicity. The platform uses census data from the US Census Bureau (for participants located in the United States) and the UK Office for National Statistics (for participants located in the United Kingdom) to divide the sample into subgroups with the same proportions as the national population. By 'national population,' we refer to the national population within the respective country of the participants, either the United States or the United Kingdom, depending on their location. The survey was published on November 11, 2021.

After applying preregistered exclusion criteria, we retained data from 470 participants for analysis (47.9 % men, 49.5 % women, 2.4 % other; $M_{age} = 31.2$, $SD_{age} = 11.4$, $Md_{education}$ = Bachelor), as shown in Fig. 1. Participants were excluded if they: (1) did not fully complete the experiment or had missing data, (2) failed an attention check to ensure careful reading, (3) completed the experiment unusually quickly—below a threshold estimated from word count and average reading speed (350 words per minute, plus a one-minute buffer), or (4) exhibited inconsistent or suspicious response patterns. Consistent with the preregistration, these criteria were applied to the initial and full sample analyses.

On average, participants indicated a political orientation score of 3.4 ($SD = 2.4$; 0 = very liberal, 10 = very conservative). The analyzed sample size was sufficiently powered with a power of 0.95 to find small to medium effects of $f = 0.18$.

### 3.2. Bias labels

We develop visualizations for biased sentences in the teaching phase using both Human-annotated and AI-generated labels that indicate whether a sentence contains biased language, as shown in Fig. 2. For Human labels, we directly use the binary sentential bias labels (Biased vs. Non-biased) from the Bias Annotations By Experts (BABE) dataset (T. Spinde, Plank et al., 2021). The BABE corpus contains 3700 sentences from diverse U.S. news outlets, balanced across topics, with bias labels annotated by five trained media experts at both the word and sentence levels. Using Krippendorff's metric α, BABE demonstrates higher inter-annotator agreement (α = 0.40) and better annotation quality compared to previous datasets (T. Spinde, Plank et al., 2021). We extract 46 sentences from BABE, evenly split between teaching and testing phases, to obtain a representative collection regarding the underlying political distribution of news outlets, as shown in Fig. 3. The political orientations of the news outlets were assigned using data from AllSides.com, which classifies news sources on a left-right scale. The 46 sentences were selected from BABE to represent diverse political orientations and biased and non-biased language (to an equal amount).

For AI labels, we first build on recent research in automated bias detection (Pryzant et al., 2020; T. Spinde, Plank et al., 2021) by training the state-of-the-art neural-based language model RoBERTa with a supervised approach on a large news corpus of biased and

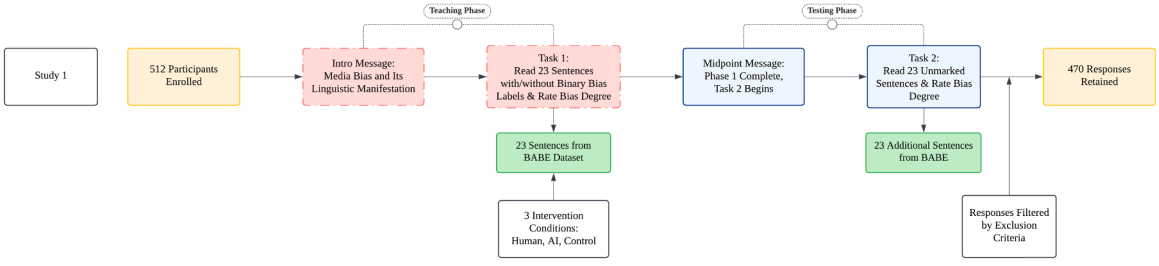


**Fig. 1. Study 1 Experimental Design:** This flowchart outlines the structure of Study 1, where 512 participants were enrolled. The teaching phase included an introduction to media bias and exposure to 23 sentences labeled by Human, AI, or Control conditions. In the testing phase, participants rated bias in 23 unmarked sentences. After applying exclusion criteria, 470 responses were retained.

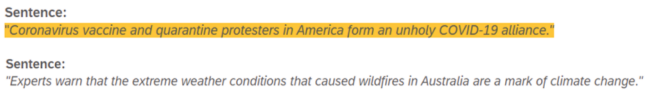


**Fig. 2. Teaching Intervention in Study 1:** An example of our sentence-level bias highlighting (above) vs. a non-biased sentence (below).

neutral text. We then use this model, BiasRoBERTa (fine-tuned on BABE and achieving a F1 score of 0.814 using 5-fold cross-validation) to classify the 23 sentences as Biased or Non-biased for our AI-generated labels. The algorithm achieved a F1 score of 0.733 for the 23 selected sentences with 5 misses and 8 false alarms compared the BABE labels. This is consistent with prior studies, which found that automated sentence-level bias detection tools are inconsistent with human raters (Horych et al., 2024). Based on the Human labels from BABE and the AI-generated labels, we create simple visualizations highlighting sentence-level bias.

### 3.3. Method

After informed consent, each participant was randomly assigned a condition and introduced to the concept of bias, and the potential intervention according to condition. In the teaching phase, participants were introduced to media bias by word choice. The participants were shown 23 teaching sentences marked according to their condition (human-annotated or AI-generated labels). However, participants were not taught about specific types of bias (e.g., linguistic, framing, or political). Instead, they were shown examples of biased sentences with neutral alternatives to help them recognize biased language in a general sense. We made this design choice aiming to focus on recognizing bias rather than teaching about the underlying mechanisms of different bias types. Subsequently, all participants were presented with 23 test sentences without any markings, regardless of condition. Finally, the participants provided general demographic information and indicated that their data could be trusted before being debriefed and compensated.

The task instructions included the statement, "In my opinion, this sentence is biased," to prompt participants to reflect on their judgment. We recognize that this phrasing could have conditioned participants to expect that the sentences were biased in some way, which may have influenced their assessments, and address the topic again in Section 5.

### 3.4. Experiment design

The experiment followed a 3-*label* (Human label/AI label/control) between × 2 *phases* (teaching/test) within design. In both the Human label and the AI label condition, teaching phase sentences were highlighted in yellow if the respective source (Human or AI) in the BABE dataset (T. Spinde, Plank et al., 2021) classified it as biased, while in the control condition and the test phase, all sentences were presented without any markings.

For each sentence, we asked participants: "Please tell us what you think about the sentence…" to assess their perception of bias.

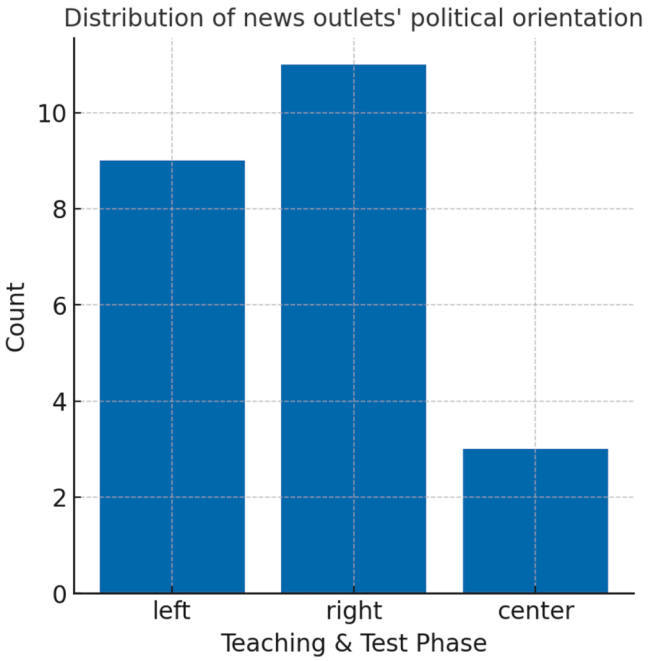


**Fig. 3. Political Distribution of News Outlets in Teaching and Testing Samples:** News outlets in the teaching and testing samples are classified based on their political orientation, as determined by AllSides.com.

Participants then responded on a 6-point Likert scale, ranging from "strongly disagree" to "strongly agree," to the statement: "In my opinion, this sentence is biased," as shown in Fig. 4. In the teaching phase, participants were presented with highlighted biased sentences and asked to rate their perceived bias. In the test phase, participants were also required to annotate bias but without any highlights, assessing their ability to generalize bias classification skills to unmarked content.

An attention check was included in both the teaching and testing phases, instructing participants to select a specific option in the bias rating question.

### 3.5. Data analysis

The bias perception rating of each sentence was binarized to compare it with the bias classifiers and obtain an accuracy score. Accordingly, the F1 score as a measure for classifying accuracy was computed for the teaching and test phase of each participant taking the human classification of the sentences according to the BABE (T. Spinde, Plank et al., 2021) dataset as the comparator. To address RQ1 and RQ2, we propose a null hypothesis for testing (We always also mention an alternative hypothesis for clarification):

**Null Hypothesis 1:** There is no difference in participants' ability to detect media bias between those taught with human labels, AI labels, and no teaching (control group).
**Alternative Hypothesis 1:** Participants taught with human labels or AI labels will exhibit a significantly higher ability to detect media bias compared to the control group that receives no teaching.

We started with analyzing descriptive statistics on the participants' labeling accuracy over all groups and survey phases. We used a 3-*label* (Human/AI/control) × 2-*phase* (teaching/test) mixed ANOVA to test how achieved labeling accuracy as F1 scores were affected by the different types of interventions using the R package afex. The ANOVA was followed by (Sidak corrected) post-hoc test using the package emmeans and a control analysis with political orientation and its interactions added to the model.

### 3.6. Results & discussion

Overall, the discovered effect of *label* ($F(467,2) = 10.36$, $p < .001$, $\eta^2_{part} = 0.042$) denotes that both the Human label ($t(467) = 4.55$, $p < .001$, $d = 0.42$) as well as the AI label ($t(467) = 2.49$, $p = .039$, $d = 0.23$) improved correct classifications of media bias compared to the control. Therefore, we reject Null Hypothesis 1, and the alternative hypothesis is considered supported based on the results of the experiment. Likewise, classifications improved from the teaching to testing phase ($F(467,1) = 32.38$, $p < .001$, $\eta^2_{part} = 0.065$; Fig. 5) across all groups (we especially address the improvement for the control group again in Section 5).

As predicted, these effects were qualified by a *label* × *phase* interaction ($F(467,2) = 5.06$, $p = .007$, $\eta^2_{part} = 0.021$): Unsurprisingly, participants with Human label intervention had better accuracy than the ones with AI label ($t(467) = 3.17$, $p = .0048$, $d = 0.29$) and controls ($t(467) = 5.32$, $p < .001$, $d = 0.49$) in the teaching phase ($F_{label\backslash teach}(467,2) = 14.34$, $p < .001$, $\eta^2_{part} = 0.06$), because their accuracy was directly evaluated by the Human labels.

While both control ($t(467) = 4.51$, $p < .001$, $d = 0.21$) and AI label ($t(467) = 4.67$, $p < .001$, $d = 0.22$) improved accuracy in the test phase and the Human label accuracy did not change systematically ($t(467) = 0.69$, $p = .4927$, $d = 0.03$), Human labels remained superior to the control in the test phase ($t(467) = 2.68$, $p = .0228$, $d = 0.25$; $F_{label\backslash test}(467,2) = 4.12$, $p = .0168$, $\eta^2_{part} = 0.02$). In the test phase, AI labels and the control could only be marginally differentiatede ($t(467) = 2.23$, $p = .0768$, $d = 0.21$) and were comparable to the Human labels ($t(467) = 0.44$, $p = .958$, $d = 0.04$).

The effectiveness of the intervention was independent of the political orientation of the taught participant (all political orientation effects: $F \leq 2.66$, all $p \geq .104$, all $\eta^2_{part} \geq 0.006$).

## 4. Study 2: effectiveness of various media literacy interventions in media bias detection

Media literacy interventions that expose participants to biased phrases or sentences effectively enhance their ability to detect bias in subsequent texts—even when annotations themselves are not entirely accurate. These effects go beyond mere exposure and attention to the problem by posing questions (Pennycook et al., 2021). However, learning effects across time also indicate that these effects of attention do exist. Moreover, interventions that highlight biased phrases without underlining the potential political motivation proved effective across the political spectrum.

Study 1 provides a proof of concept, demonstrating that media bias interventions can yield small to moderate learning effects. However, it was limited to single sentences and basic feedback learning. Study 2 expands this approach to bias detection at the article level, as illustrated in Fig. 6, raising key questions about optimizing media literacy interventions for teaching: Can interventions

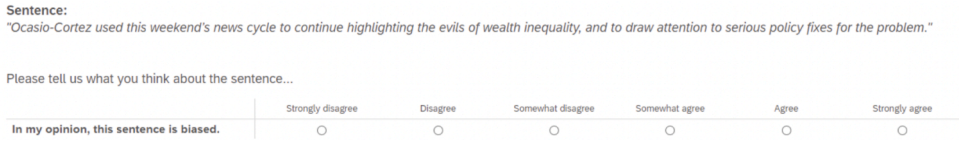


**Fig. 4. Sample Survey Question on Media Bias Perception:** A survey question presenting a sentence and asking participants to assess media bias.

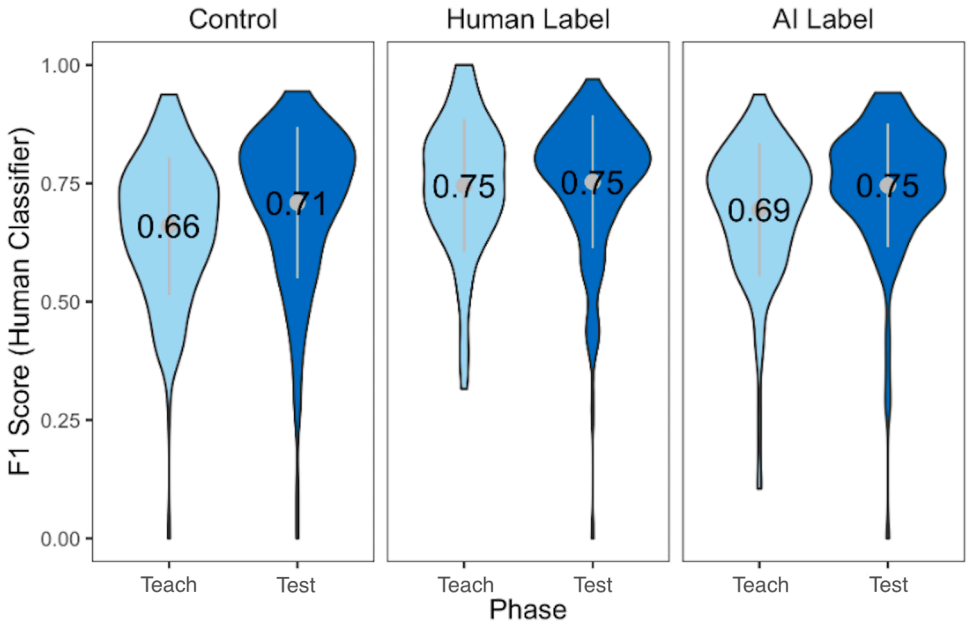


**Fig. 5. Distribution of Participants' F1 Scores:** F1 score distribution across all survey groups and phases.

remain effective when applied to full-length texts rather than isolated sentences? Is flagging entire biased sentences sufficient, or do consumers benefit more from identifying specific biased words and receiving neutral rewording suggestions? Additionally, does providing political context for an article improve bias detection by illustrating how certain phrases are used in a specific political milieu (S. D'Alonzo & Tegmark, 2022)?

### 4.1. Participants

The study was sampled on the sampling platform Prolific. Qualifications entailed equal gender distribution, USA residency, fluency in English, a minimum of 50 and a maximum of 10,000 completed tasks, and an approval rate of >95 % via Prolific settings. Sampling occurred on January 15, 2022, and workers were compensated with 2.25£. A total of 1121 participants were recruited, of which $N_{full} =$ 846 participants (434 female, 391 male, 16 other, 5 not stated; age: $M = 35.15$, $SD = 13.28$) provided data used in the analyses. Same as in Study 1, we removed 275 participants based on the following exclusion criteria: (1) Participants who did not fully complete the experiment or were missing data; (2) Participants who failed an attention check designed to ensure that the participants were carefully reading the content; (3) Participants who completed the experiment in an unusually short amount of time, below an estimated threshold based on word count and average reading speed (350 words per minute, with an additional one-minute buffer); and (4) Participants flagged as unreliable based on inconsistent or suspicious patterns of responses. These criteria were consistent with the preregistered exclusions and applied to both the initial and full sample analyses. On average, participants indicated a political orientation score of −17.54 ($SD = 27.30$; −50 = very liberal, 50 = very conservative). This sample is 95 % powered to detect small effects of at least $f = 0.13$ for the preregistered comparisons.

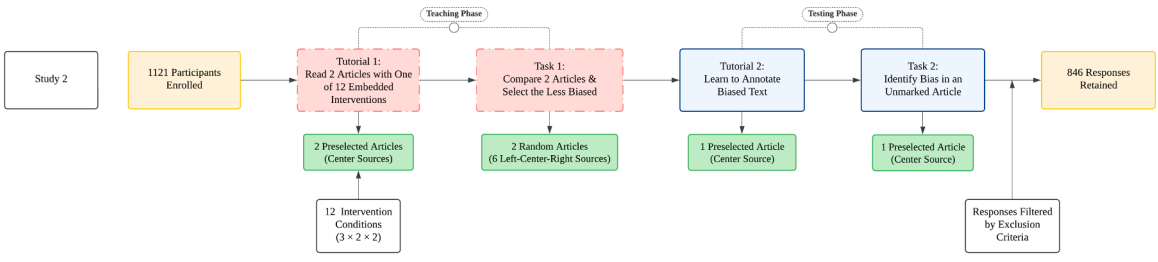


**Fig. 6. Study 2 Experimental Design:** This flowchart illustrates Study 2, which enrolled 1121 participants. The teaching phase consisted of a tutorial with two preselected articles embedded with one of 12 intervention conditions, followed by a task comparing two random articles from left-center-right sources. In the testing phase, participants learned to annotate biased text before identifying bias in an unmarked article. After filtering, 846 responses were retained.

### 4.2. Method

After providing informed consent and completing demographic questions, participants indicated their political orientation using a visual slider ranging from –50 (very liberal) to 50 (very conservative). They were then assigned to one of twelve media literacy intervention conditions and completed the study in two phases, as illustrated in Fig. 6.

In the teaching phase, participants first received a tutorial introducing the visualization interface based on their assigned condition. Fig. 7 illustrates an intervention condition where biased phrase highlighting, advanced biased sentence labeling, and discriminative phrase detection were all present. Participants then engaged with interventions visualizing biased language in two articles from center sources on the same topic on the same topic. Next, they read two biased articles, quasi-randomly selected from a pool of six (two each from left, center, and right sources), and determined which was less biased.

In the testing phase, all participants first completed a tutorial on marking biased language by clicking on sentences and phrases they deemed biased. They then read an unmarked news article and annotated all biased text. Finally, participants were debriefed and compensated.

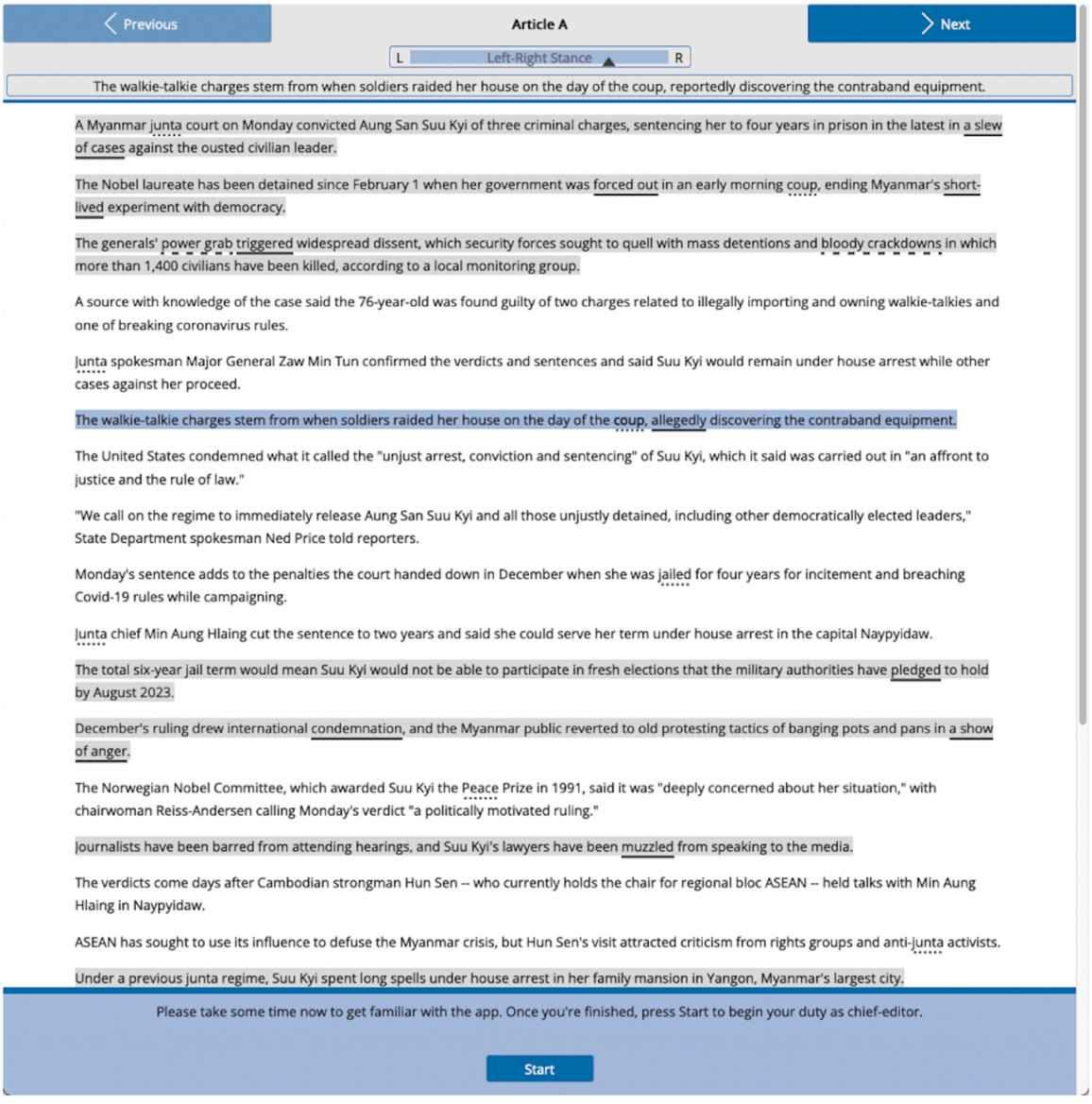


**Fig. 7. Screenshot of the teaching tutorial:** An example featuring interventions for biased phrases, advanced biased sentences, and discriminative phrase detection. The selected blue sentence contains a highlighted discriminative phrase ("coup"). At the top, the "Analysis Bar" is displayed, showing the left-right scale and the "Synonyms Field". Instructions were provided throughout the experiment via the blue overlay at the bottom.

### 4.3. Materials and design

The intervention conditions in the teaching phase followed a 3-*biased sentence labeling* (absent/simple/advanced) × 2 *biased phrase labeling* (absent/present) × 2 *politicized phrase labeling* (absent/present) between design. Fig. 7 shows the task screen entailing the all-label conditions. For each article, politicized phrases were obtained from S. topic-related discriminative phrase categorization and min-max normalized per topic, biased phrases and sentences were determined by two raters agreeing on the classification following the annotation guidelines by T. .

In the label conditions, biased phrases were underlined, politicized phrases were highlighted with a dotted underline, and the combination of both with a dashed underline. Biased sentences were highlighted in the respective conditions with a grey background color. Additional information (i.e., the left-right bias of politicized phrases and/or a synonymous unbiased sentence) was displayed above the article in the so-called "Analysis Bar" for conditions entailing these advanced communications when selected by a mouse click (politicized phrases condition, advanced biased sentences condition).

For both phases, articles about the same topic and political ratings were collected using GroundNews. Table 1 presents the length and proportion of biased words for each article used in the study, including their processed versions. Following findings that the congruence of the political attitude of an article and a participant may affect its perception (Spinde et al., 2022), the articles for the teaching phase were selected across political isles. Conversely, the test article was chosen to be neutral and uncontroversial to provide similar preconditions for all participants, to investigate whether participants could generalize their bias detection skills beyond clearly biased content. We based the choice of the neutral test article on existing findings, showing that neutral articles in testing can reduce heuristic bias identification and motivate reasoning effects (Kahan, 2013; Pennycook et al., 2020).

Accordingly, the tutorial article included two biased articles about Myanmar's former president Aung San Suu Kyi. The teaching task consisted of two biased articles about weapon restrictions in the USA, specifically the trial of Kyle Rittenhouse, sampled from two left, center, and right media sources. To grant comparable stimulus intervention conditions, all articles were algorithmically modified based on the BABE dataset (T. Spinde, Plank et al., 2021) to contain 3.33 %, 6.66 %, or 10 % of biased words, respectively. As shown in Table 1, the percentages aim to reflect a structured bias exposure with low, medium, and high bias conditions (based on the number of biased phrases found in the original articles in BABE), facilitating systematic comparisons of participants' responses to varying bias levels. In addition, longer articles were manually trimmed with the lowest loss of information possible to contain roughly the same number of words. A clustered random assignment to the articles ensured that one article was presented in a less biased version than the other. The test phase tutorial used an article about the shutdown of an independent Hong Kong news portal. In the test trial, all participants annotated an article about the launch of the James Webb telescope. For detailed information about the articles, see the materials at osf.io.

In the teaching phase, participants had to determine which of the two presented articles they deemed more biased. In the test phase, participants were asked to indicate biased phrases in the test article by highlighting them with a double click.

### 4.4. Data analysis

The F1 score was computed for the test article on sentence level. The classification of biased sentences was obtained from T. Spinde, Plank et al. (2021). Accordingly, marked biased sentences were counted as hits, marked unbiased sentences as false alarms, unmarked biased sentences as misses, and unmarked unbiased sentences as correct rejections. Please note that F1 scores are lower than in Study 1 because participants were not forced to react to all sentences yielding a higher rate of misses.

To address RQ2 and RQ3, we propose the following two null hypotheses for testing:

**Null Hypothesis 2:** All intervention types (e.g., sentence-level vs. phrase-level bias highlighting) show no difference in their effectiveness in teaching participants to detect biased language.

**Alternative Hypothesis 2:** Some intervention types (e.g., phrase-level highlighting) are significantly more effective in teaching participants to detect biased language than others

(*continued on next page*)

**Table 1**
Number of words and proportion of biased words for every used article, including processed versions.

| | | | Number of words / % of biased words | | | |
|---|---|---|---|---|---|---|
| Phase | News Source | Bias | Initially | Low | Medium | High |
| **Tutorial 1** | AFP News | Center | 643/5.49 | | | |
| | Coote and Maresca | Center | 621/6.48 | | | |
| **Task 1** | Lamoureux | Left | 671/9.86 | 537/3.37 | 543/6.66 | 545/9.93 |
| | Lutz | Left | 719/16.23 | 520/3.45 | 530/6.71 | 533/10 |
| | Tarm and Bauer | Center | 852/9.69 | 585/3.44 | 582/6.66 | 587/10 |
| | Tarm et al. | Center | 418/11.69 | 415/3.33 | 414/6.66 | 414/9.96 |
| | Shiver | Right | 552/12.61 | 523/3.31 | 539/6.69 | 544/10.06 |
| | Lane | Right | 445/12.60 | 436/3.46 | 433/6.59 | 446/10.04 |
| **Tutorial 2** | Cheng et al. | Center | 215/5.97 | | | |
| **Task 2** | Amos | Center | 626/9.25 | | | |

(*continued*)

**Null Hypothesis 3:** Participants' backgrounds, such as political inclinations, do not interact with different visualizations and do not jointly influence their learning outcomes.
**Alternative Hypothesis 3:** Participants' backgrounds, such as political inclinations, interact with different visualizations and jointly influence their learning outcomes.

To test how the feedback affected the ability to correctly identify the more biased article of the two presented in the teaching phase, a 3-*biased sentence labeling* (absent/simple/advanced) × 2-*biased phrase labeling* (absent/present) × 2-*politicized phrase labeling* (absent/present) logistic regression with the article choice (higher 0/lower biased article 1) was performed as the criterion and effect-coded factors.

A 3 *biased sentence labeling* (absent/simple/advanced) × 2-*biased phrase labeling* (absent/present) × 2 *politicized phrase labeling* (absent/present) between-subject ANOVA on the extracted F1 scores was performed to test whether the feedback affected participant's ability to identify biased sentences. Effects of *biased sentence labels* were followed by simple contrast (Bonferroni-corrected) and interactions by simple effects analyses using the R package emmeans.

To test how political orientation affected the feedback interventions, both designs were extended to GLMs, adding political orientation (scaled from −50 (left) to +50 (right)), article orientations, and all interactions. Article orientations were formed as an average of the orientation of both presented articles in the teaching phase coded −1 for left, 0 for neutral, and 1 for right articles. All factors were effect coded, and all interactions were included.

The analyses were preregistered under OSF.IO/3VDCH. Due to a technical redirection error after experiment completion, data of about three times the preregistered sample size was accidentally acquired. Analyses reported use all the data using the preregistered screening criteria and additionally excluding one individual taking implausibly long (>9 SDs longer than the mean), as the data of the targeted sample are indiscernible from the oversampled one. With this effective sample size of 846, we are powered with a 0.95

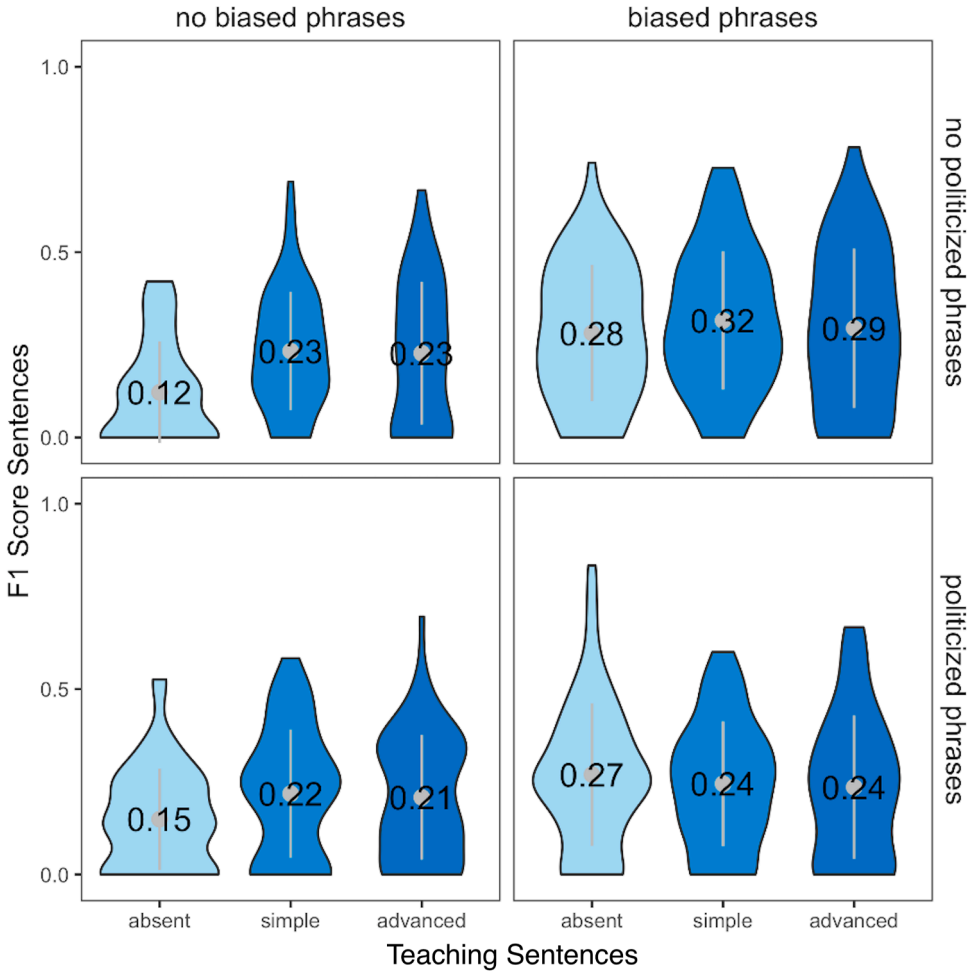


**Fig. 8.** Distribution of Participants' F1 Scores in the Testing Phase: F1 score distribution across all survey groups during the testing phase.Biased sentence detection in the test phase.

probability to find significant small effects of at least $f = 0.13$.

### 4.5. Results & discussion

#### 4.5.1. Biased article selection in the teaching phase

Overall, 563 participants (66.6 %) correctly identified the more biased article with a range in the experimental group from 53.7 % in the all-absent control group to 78.8 % for the advanced sentence feedback with politicized, but no biased phrases group. Yet regarding experimental effects, participants only correctly identified the higher biased article of the tutorial set more consistently when being provided with *politicized phrase* labeling ($\chi^2(1) = 6.19$, $p = .0129$, $OR = 1.46$; 62.7 % correct without vs. 70.8 % with politicized phrases) with no other effect being significant (all $p \geq .086$). This illustrates that participants were directly affected in their ranking decision by the political feedback of the experimenter on the degree of bias.

This effect remained significant for neutral articles and political orientation ($\chi^2(1) = 9.51$, $p = .002$), but was qualified by a higher order interaction with both article and participant orientation and *biased sentence* labeling ($\chi^2(1) = 13.80$, $p = .001$): When articles inconsistent with a participant view had to be rated with no bias sentence labeling, *politicized phras*e labeling was rather reducing the accuracy of the decision.

Overall, participants were better able to correctly identify biased sentences in the test phase, when they were taught with a *biased sentence* ($F(834,2) = 5.58$, $p = .004$, $\eta^2_{part} = 0.012$) and *biased phrase* labels ($F(834,1) = 44.00$, $p < .001$, $\eta^2_{part} = 0.048$). However, F1-scores were decreased by *politicized phrase* labels ($F(834,1) = 4.16$, $p = .042$, $\eta^2_{part} = 0.005$; see Fig. 8). While a negative effect of politicized phrases in this context is somewhat surprising and novel, it is in line with earlier findings indicating that the political dimension is somewhat unrelated and may even obscure the ability to detect biases (S. D'Alonzo & Tegmark, 2022) – maybe by reminding the reader of their own political identity and the political dimension of this task.

The interaction of both successful interventions (*biased sentence* × *biased phrase:* $F(834,2) = 4.16$, $p = .002$, $\eta^2_{part} = 0.01$) further illustrated that *biased phrase* labelling was sufficient to evoke effects: Simple and advanced biased sentence labels in the teaching phase were only superior in F1 accuracy compared to control, when no additional biased phrase labels were presented ($t_{simple \text{ vs. } absent\backslash no\ biased\ phrases}(834) = 4.28$, $p < .001$, $d = 0.51$; $t_{advanced \text{ vs. } absent\backslash no\ biased\ phrases}(834) = 4.28$, $p < .001$, $d = 0.47$; $t_{simple \text{ vs. } absent\backslash biased\ phrases}(834) = 0.22$, $p = 1$, $d = 0.03$; $t_{advanced \text{ vs. } absent\backslash biased\ phrases}(834) = 0.50$, $p = 1$, $d = 0.06$).

On the other hand, *biased phrase* labeling always increased performance regardless of *biased sentence* labeling (all $t(834) \geq 2.22$, all $p \geq .027$, all $d \geq 0.26$). In addition, there seemed to be a marginal trend that *biased phrase* labeling was to some degree less effective when the intervention also contained *politicized phrase* labels ($F(834,1) = 3.42$, $p = .065$, $\eta^2_{part} = 0.004$). No other effect on sentence-level F1-scores emerged (all $p \geq .167$, all $\eta^2_{part} \leq 0.004$). Overall, the results suggest that we reject Null Hypothesis 2, and the alternative hypothesis is supported, as different visualization strategies demonstrated varying effectiveness in teaching participants to detect media bias.

Testing for potential mitigation by political views of both the interventions and the participants, we found that effects of *biased phrase* labels in the teaching phase hold controlling for political ideologies ($F(798,1) = 29.100$, $p < .001$, $\eta^2_{part} = 0.033$). Still, more conservative participants were generally less accurate ($F(798,1) = 6.48$, $p = .011$, $\eta^2_{part} = 0.007$) and also decreased the effects of *biased sentence* labeling (*biased sentence* × *political affiliation:* $F(798,2) = 5.43$, $p = .005$, $\eta^2_{part} = 0.01$), so that the effects were not significantly present in politically neutral participants (*biased sentence* $F(798,2) = 1.04$, $p = .353$, $\eta^2_{part} = 0.002$). This indicates that participants' political inclinations did interact with the visualizations and affect learning outcomes, particularly for conservative participants, suggesting that we fail to reject Null Hypothesis 3 and support the alternative hypothesis. Politicized phrase labels may have introduced additional cognitive load or motivated reasoning effects, which is a promising perspective for future work and further investigation, as we detail in Section 5. There were no further effects of political viewpoints (all $p \geq .168$, all $\eta^2_{part} < 0.01$).

## 5. General discussion

Media literacy interventions that expose participants to media-biased phrases or sentences have consistently demonstrated efficacy in enhancing the ability to detect biases in subsequent texts—even when the annotations themselves might not be entirely accurate. Notably, marking specific phrases as biased, without emphasizing potential political motivations, proves effective across the political spectrum. The results indicate that, although media bias is a complex concept, visualizations highlighting phrasal- and sentential-level biased content act as a nudge, encouraging readers to think more critically about bias in the news. While pinpointing every biased instance within a comprehensive text can be challenging—possibly due to diminished motivation to identify all instances—accuracy at the sentence level shows an increase by up to an F1-Score of 0.08, signifying a medium-sized effect. Our study shows that while automated labels are not as accurate or effective as Human labels, they can still contribute to raising news consumers' awareness of media bias to some extent. Automated labels can be feasibly scaled to handle the vast amount of online content, making them especially useful in cases where human annotations are unavailable. It's imperative to clarify that the observed learning effects aren't merely a byproduct of heightened attention to bias caused by the tutorial on bias annotation. Control groups in Study 1 are also provided with identical instructions to rate sentences and articles, yet the learning effect observed in this group suggests that minimal exposure effects coexist alongside the more pronounced learning effects of feedback labels. This is also visible in the F1-scores, where we see a significant improvement in the control groups. While the exact reason here is unclear, the effect may be due to selective attention, where participants in the intervention groups focused on highlighted bias rather than developing a broader recognition strategy. Also, the control group engaged in implicit learning (Kolb, 1984), potentially improving bias detection through task engagement without explicit guidance. Future research should explore whether bias annotations, in general, shape recall patterns.

While we found that at the sentence level, learning effects can be generalized to new, unannotated news materials with unseen topics to detect biased language, the annotations do not consistently result in the accurate classification of entire articles. This discrepancy might stem from a less defined notion of bias at the article level, allowing political inclinations to exert a more pronounced influence when comparing articles. This theory aligns with the lower performance of politicized language highlighting, consistent with prior findings on the impact of visualizing political affiliations (Spinde et al., 2022). In future work building upon our study, we aim to incorporate post-task interviews or surveys to assess whether participants perceived politicized labels as distracting from broader bias perception. Additionally, it seems promising to examine how the results generalize across different journalistic styles and publication biases. Our current study is limited in assessing its generalizability across different outlet types, but also across cultures and languages. Future research should test whether bias detection performance differs when applied to news from explicitly partisan sources versus those aiming for neutrality and vary in selected outlets, worldwide. Understanding how interventions interact with different journalistic norms could further refine bias literacy education and its real-world applicability. Additionally, the generalizability of the results is limited by the temporal scope of the study. Since the articles used in the study are restricted to specific topics or isolated sentences, the generalizability of the learning effects to a broader set of articles might be constrained. This limitation is particularly relevant given the evolving nature of media and journalistic practices, as well as shifts in political discourse over time. Future work could focus on how these interventions perform as media landscapes change, as well as how different political climates might influence bias detection accuracy and effectiveness over time.

Given that the articles used in the study are restricted to specific topics or isolated sentences, the generalizability of the learning effects to a broader set of articles might be constrained. Addressing this limitation is challenging due to the practical constraints of sample size—a challenge shared by many studies in the media bias domain. Another area of concern is the longevity of the learning effects, which the current study design did not cover.

Regarding the participants' prior knowledge of biases, it is unclear whether any conditioning effect of the tutorial may also play a role in their performance. For example, the phrasing of the instruction, "In my opinion, this sentence is biased," could have led participants to be more inclined to perceive all sentences as biased to some extent, which might have influenced their judgments. While we focused on teaching participants to recognize biased language, participants' knowledge of biases could have affected their performance, which also points to an area for future research, where we could investigate how different levels of media bias literacy influence participants' ability to detect bias, as well as explore whether the use of neutral, non-conditioning phrases might lead to different outcomes. In general, how to ask about bias is a major factor in media bias research, and results on the impact of questions and prior knowledge are scarce (Spinde, Kreuter et al., 2021). Since the highlighting of biased language depends on context and can only be scaled through AI-driven detection, further research should focus on improving automated highlighting to ensure its accuracy in identifying bias in new news content. Even more, after demonstrating the effectiveness of our approach, the next question is how to build applications that provide these benefits for news consumers. One challenge in this regard is the potential for dataset shifts, where the writing styles and political framing in news articles evolve over time. Models trained on older datasets may struggle to detect bias in newer content with different stylistic trends. However, we believe this challenge to be decreasing since large language models are constantly improving in their adaptability and generalization abilities (Horych et al., 2025; OpenAI, 2024), and it is constantly becoming easier to also train and re-train dedicated models for specific aspects, languages, or datasets.

Our findings hold practical implications for developing news-reading platforms that offer bias indicators for everyday news consumption and for use in schools and media literacy education. Recent changes in social media, particularly the removal of third-party moderators as of January 2025 (Duffy, 2025), make bias-aware news consumption and even more interesting field. As moderation shifts, other tools will be even more critical for guiding news consumers in identifying bias without external oversight. As citizens with diverse perspectives on societal issues help strengthen democracy, this research can serve as a first step toward developing strategies, tools, and resources for promoting media literacy and bias-aware news consumption. However, further research is needed to explore how these insights can be effectively applied to real-world news consumers and integrated into educational content. Even more, our current research only showed how viable automatically created models can generally be, and we focus on whether there is an effect in general. Since new models emerge regularly and continuously improve bias classifications, our results indicate that investigating model differences in similar tasks in detail is a promising direction for future work.

## 6. Conclusion

This study addresses the gap in systematically assessing the generalizability of media bias interventions by conducting two experiments to evaluate participants' ability to identify bias in plain news articles on new topics after learning with various bias labels. Our findings demonstrate that both Human and AI-generated bias labels can effectively enhance media bias awareness, and that phrasal labeling proves to be the most effective visualization, regardless of political tendencies. Notably, highlighting politicized language and inclinations may have negative effects, aligning with previous findings.

Compared to prior research, this study uniquely examines the differential effectiveness of Human and AI-generated labels and various bias-indicators across varied intervention conditions, offering insights into optimal labeling strategies. Theoretical implications include proposing empirical frameworks for evaluating the generalizability of bias-awareness intervention to new contexts, providing a systematic approach to measure learning effects across diverse content.

Practically, the findings suggest the potential of integrating bias indicators into news-reading platforms and media literacy curricula to enhance consumers' ability to detect bias. A key gap to address next is translating these findings into real-world applications, such as developing tools for bias-aware news consumption or incorporating bias-awareness interventions into educational programs. Future research should also refine the accuracy of AI-generated labels when applied to new content, ensuring they adapt to

the dynamic nature of news and consistently raise readers' awareness of bias.

## Funding

This work was supported by the Hanns-Seidel Foundation, the German Academic Exchange Service (DAAD), and the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022). It was partially supported by JSPS KAKENHI Grants JP21H04907 and JP24H00732, by JST CREST Grant JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, and by JST K Program Grant JPMJKP24C2 Japan. It was also funded by the German Federal Ministry of Education and Research (BMBF) through the DAAD (German Academic Exchange Service) and the German Research Foundation [DFG] under Grant 441541975. None of the funders played any role in the study design or publication-related decisions.

## Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to proofread the document. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

## CRediT authorship contribution statement

**Timo Spinde:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **Fei Wu:** Writing – review & editing, Writing – original draft. **Wolfgang Gaissmaier:** Funding acquisition. **Gianluca Demartini:** Writing – review & editing, Funding acquisition. **Isao Echizen:** Writing – review & editing. **Helge Giese:** Writing – original draft, Methodology, Conceptualization.

## Acknowledgments

We would like to thank Smi Hinterreiter, Anna Bahß, David Krieger, and Tilman Hornung for their efforts in organizing data assessment for the studies and Bela Gipp for his scholarly advice.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2025.104244.

## Data availability

All research data are published within the paper and respective repositories.

## References

Auxier, B. E., & Vitak, J. (2019). Factors motivating customization and echo chamber creation within digital news environments. *Social Media + Society, 5*(2), Article 205630511984750. https://doi.org/10.1177/2056305119847506

Bernhardt, D., Krasa, S., & Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics, 92*(5), 1092–1104. https://doi.org/10.1016/j.jpubeco.2008.01.006

Beukeboom, C. J., & Burgers, C. (2017-07). *"Linguistic Bias". In: Oxford Research Encyclope- dia of Communication.* Oxford University Press. isbn: 978-0-19-022861-3 https://doi.org/10.1093/acrefore/9780190228613.013.439.

D'Alonzo, S., & Tegmark, M. (2022). Machine-learning media bias. *PLOS ONE, 17*(8), Article e0271947. https://doi.org/10.1371/journal.pone.0271947

Dhanani, L. Y., & Franz, B. (2020). The role of news consumption and trust in public health leadership in shaping COVID-19 knowledge and prejudice. *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.560828

Druckman, J. N., & Parkin, M. (2005). The impact of Media bias: How editorial slant affects voters. *The Journal of Politics, 67*(4), 1030–1049. https://doi.org/10.1111/j.1468-2508.2005.00349.x

Duffy, C. (2025). Meta gets rid of fact checkers and says it will reduce 'censorship' | CNN Business. *CNN*. https://www.cnn.com/2025/01/07/tech/meta-censorship-moderation/index.html.

Eberl, J.-M., Boomgaarden, H. G., & Wagner, M. (2017). One bias fits all? Three types of Media bias and their effects on party preferences. *Communication Research, 44*(8), 1125–1148. https://doi.org/10.1177/0093650215614364

Goodsett, M. (2023). Applying misinformation interventions to library instruction and outreach. *Journal of New Librarianship, 8*, 78. https://heinonline.org/HOL/Page?handle=hein.journals/jnwlibsh8&id=210&div=&collection=.

Goodsett, M. (2024). Chaos creator: Misinformation inoculation in information literacy instruction. *Journal of Information Literacy, 18*(2), 56–86. https://doi.org/10.11645/18.2.13

Hamborg, F., Donnay, K., & Gipp, B. (2019). Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries, 20*(4), 391–415. https://doi.org/10.1007/s00799-018-0261-y

Hameleers, M. (2022). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society, 25*(1), 110–126. https://doi.org/10.1080/1369118X.2020.1764603

Happer, C., & Philo, G. (2013). The role of the Media in the construction of public belief and social change. *Journal of Social and Political Psychology, 1*(1), 321–336. https://doi.org/10.5964/jspp.v1i1.96

He, Z., Majumder, B. P., & McAuley, J. (2021). Detect and perturb: Neutral rewriting of biased and sensitive text via gradient-based decoding. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 4173–4181). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.352.

Henderson, L., & Green, C. (2020). Making sense of microplastics? Public understandings of plastic pollution. *Marine Pollution Bulletin, 152*, Article 110908. https://doi.org/10.1016/j.marpolbul.2020.110908

Hinterreiter, S., Wessel, M., Schliski, F., Echizen, I., Latoschik, M. E., & Spinde, T. (2025). *NewsUnfold: Creating a news-reading application that indicates linguistic media bias and collects feedback, 19*. AAAI https://media-bias-research.org/wp-content/uploads/2024/07/Preprint_ICWSM_25_NewsUnfold.pdf.

Hinterreiter, S., Spinde, T., Oberdörfer, S., Echizen, I., & Latoschik, M. E. (2024). News Ninja: Gamified annotation of linguistic bias in online news. *Proceedings of the ACM on Human-Computer Interaction, 8 (CHI PLAY). Association for Computing Machinery*. https://doi.org/10.1145/3677092.

Horych, T., Wessel, M. P., Wahle, J. P., Ruas, T., Waßmuth, J., Greiner-Petter, A., Aizawa, A., Gipp, B., & Spinde, T. (2024). MAGPIE: Multi-task analysis of Media-bias generalization with pre-trained identification of expressions. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)* (pp. 10903–10920). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.952/.

Horych, T., Mandl, C., Ruas, T., Greiner-Petter, A., Gipp, B., Aizawa, A., & Spinde, T. (in press). The promises and pitfalls of LLM annotations in dataset labeling: A case study on media bias detection. *Findings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics (NAACL 2025)*. Association for Computational Linguistics.

Hube, C., & Fetahu, B. (2019). Neural based statement classification for biased language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 195–203). https://doi.org/10.1145/3289600.3291018

Iandoli, L., Primario, S., & Zollo, G. (2021). The impact of group polarization on the quality of online debate in social media: A systematic literature review. *Technological Forecasting and Social Change, 170*, Article 120924. https://doi.org/10.1016/j.techfore.2021.120924

Islam, R. (2008). *Information and public choice: From media markets to policymaking*. World Bank Publications.

Joris, G., Vercoutere, S., De Clercq, O., Van Damme, K., Mechant, P., & De Marez, L. (2024). Nudging towards exposure diversity: Examining the effects of news recommender design on audiences' News exposure behaviours and perceptions. *Digital Journalism, 12*(8), 1118–1139. https://doi.org/10.1080/21670811.2022.2106445

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making, 8*(4), 407–424. https://doi.org/10.1017/S1930297500005271

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Prentice Hall.

Lee, N., Bang, Y., Yu, T., Madotto, A., & Fung, P. (2022). NeuS: Neutral multi-news summarization for mitigating framing bias. In M. Carpuat, M.-C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 3131–3148). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.228.

Lei, Y., Huang, R., Wang, L., & Beauchamp, N. (2022). Sentence-level Media bias analysis informed by discourse structures. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 10040–10050). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.682.

Liu, R., Wang, L., Jia, C., & Vosoughi, S. (2021). Political depolarization of news articles using attribute-aware word embeddings. *Proceedings of the International AAAI Conference on Web and Social Media, 15*, 385–396. https://doi.org/10.1609/icwsm.v15i1.18069

Lu, C., Hu, B., Bao, M.-M., Wang, C., Bi, C., & Ju, X.-D. (2024). Can Media literacy intervention improve fake news credibility assessment? A meta-analysis. *Cyberpsychology, Behavior, and Social Networking, 27*(4), 240–252. https://doi.org/10.1089/cyber.2023.0324

Mattis, N., Masur, P., Möller, J., & van Atteveldt, W. (2024). Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *New Media & Society, 26*(7), 3681–3706. https://doi.org/10.1177/14614448221104413

McCarthy, K. J., & Dolfsma, W. (2014). Neutral Media? Evidence of Media bias and its economic impact. *Review of Social Economy, 72*(1), 42–54. https://doi.org/10.1080/00346764.2013.806110

Munson, S., Lee, S., & Resnick, P. (2021). Encouraging reading of diverse political viewpoints with a browser widget. *Proceedings of the International AAAI Conference on Web and Social Media, 7*(1), 419–428. https://doi.org/10.1609/icwsm.v7i1.14429

OpenAI. (2024). GPT-4 technical report. *arXiv*. https://doi.org/10.48550/arXiv.2303.08774

Paramita, M. L., Kasinidou, M., Kleanthous, S., Rosso, P., Kuflik, T., & Hopfgartner, F. (2024). Towards improving user awareness of search engine biases: A participatory design approach. *Journal of the Association for Information Science and Technology, 75*(5), 581–599. https://doi.org/10.1002/asi.24826

Park, S., Kang, S., Chung, S., & Song, J. (2009). NewsCube: Delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443–452). https://doi.org/10.1145/1518701.1518772

Park, S., Kang, S., Chung, S., & Song, J. (2012). A computational framework for Media bias mitigation. *ACM Trans. Interact. Intell. Syst., 2*(2), 8. https://doi.org/10.1145/2209310.2209311, 1–8:32.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on Social Media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Piksa, M., Noworyta, K., Gundersen, A., Kunst, J., Morzy, M., Piasecki, J., & Rygula, R. (2024). The impact of confirmation bias awareness on mitigating susceptibility to misinformation. *Frontiers in Public Health, 12*. https://doi.org/10.3389/fpubh.2024.1414864

Pryzant, R., Martinez, R. D., Dass, N., Kurohashi, S., Jurafsky, D., & Yang, D. (2020). Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence, 34*(01), 480–489. https://doi.org/10.1609/aaai.v34i01.5385

Rieger, A., Draws, T., Theune, M., & Tintarev, N. (2024). Nudges to mitigate confirmation bias during web search on debated topics: Support vs. Manipulation. *ACM Transactions on the Web, 18*(2), 1–27. https://doi.org/10.1145/3635034

Spinde, T., Hamborg, F., Donnay, K., Becerra, A., & Gipp, B. (2020). Enabling news consumers to view and understand biased news coverage: A study on the perception and visualization of Media bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 389–392). https://doi.org/10.1145/3383583.3398619

Spinde, T., Rudnitckaia, L., Kanishka, S., Hamborg, F., Bela, G., & Donnay, K. (2021). MBIC – A media bias annotation dataset including annotator characteristics. *In Proceedings of the iConference 2021*. https://doi.org/10.6084/m9.figshare.17192924

Spinde, T., Kreuter, C., Gaissmaier, W., Hamborg, F., Gipp, B., & Giese, H. (2021a). Do you think it's biased? How to ask for the perception of Media bias. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 61–69). https://doi.org/10.1109/JCDL52503.2021.00018

Spinde, T., Plank, M., Krieger, J. D., Ruas, T., Gipp, B., & Aizawa, A. (2021b). Neural Media bias detection using distant supervision with BABE - Bias annotations by experts. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 1166–1177. https://doi.org/10.18653/v1/2021.findings-emnlp.101

Spinde, T., Rudnitckaia, L., Mitrović, J., Hamborg, F., Granitzer, M., Gipp, B., & Donnay, K. (2021c). Automated identification of bias inducing words in news articles using linguistic and context-oriented features. *Information Processing & Management, 58*(3), Article 102505. https://doi.org/10.1016/j.ipm.2021.102505

Spinde, T., Jeggle, C., Haupt, M., Gaissmaier, W., & Giese, H. (2022). How do we raise media bias awareness effectively? Effects of visualizations to communicate bias. *PLOS ONE, 17*(4), Article e0266204. https://doi.org/10.1371/journal.pone.0266204

Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., & Gipp, B. (2024). The Media bias taxonomy: A systematic literature review on the forms and automated detection of Media bias. *arXiv*. http://arxiv.org/abs/2312.16148.

Spinde, T. (2021). An interdisciplinary approach for the automated detection and visualization of Media bias in news articles. In *2021 International Conference on Data Mining Workshops (ICDMW)* (pp. 1096–1103). https://doi.org/10.1109/ICDMW53433.2021.00144

Sugimoto, A., Nomura, S., Tsubokura, M., Matsumura, T., Muto, K., Sato, M., & Gilmour, S. (2013). The relationship between Media consumption and health-related anxieties after the Fukushima Daiichi nuclear disaster. *PLOS ONE, 8*(8), Article e65331. https://doi.org/10.1371/journal.pone.0065331

Terren, L. T. L., & Borge-Bravo, R. B.-B. R. (2021). Echo chambers on social media: A systematic review of the literature. *Review of Communication Research, 9*. https://lex-localis.org.

Van Der Meer, T. G. L. A., & Hameleers, M. (2021). Fighting biased news diets: Using news media literacy interventions to stimulate online cross-cutting media exposure patterns. *New Media & Society, 23*(11), 3156–3178. https://doi.org/10.1177/1461444820946455

Wessel, M., Horych, T., Ruas, T., Aizawa, A., Gipp, B., & Spinde, T. (2023). Introducing MBIB - The first Media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2765–2774). https://doi.org/10.1145/3539618.3591882

Williams, A. (1975). Unbiased study of television news bias. *Journal of Communication, 25*(4), 190–199. https://doi.org/10.1111/j.1460-2466.1975.tb00656.x