Battle Misinformation: Detecting Cognitive Bias with Eye Movements

XIUGE CHEN, The University of Melbourne, Australia
BENJAMIN TAG, The University of Melbourne, Australia
CATHERINE E. DAVEY, The University of Melbourne, Australia
TILMAN DINGLER, The University of Melbourne, Australia

To cope with the ever-increasing amount of information, people take shortcuts to make quick judgements and avoid high cognitive load. These cognitive shortcuts, however, can lead to cognitive biases, which affect our perception and lead to subjective realities that potentially result in poor judgement and decision-making. Traditionally, cognitive biases are detect using self-assessment tools, which are both subjective and prone to falsification. To better quantify and mitigate the impact of biases, more objective and unobtrusive measurements are required. In this work, we use eye movement features to detect confirmation bias and cognitive dissonance, two common biases in the context of online news consumption. We presented an approach to systematically induce and elicit these two biases and train a machine-learning model that can detect whether news in the form of images or text contradict or align with a user's view. We conducted a study with 33 participants achieving up to 68% accuracy with a general model. We were able to boost its accuracy up to 76% when focusing on particular, polarizing topics. Our models outperform existing solutions in terms of both performance and generalizability, and can be used to study the occurrence of cognitive biases in online news media to devise strategies for introspection and mitigation of their effects.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Empirical studies in HCI*; *Ubiquitous and mobile computing systems and tools.*

Additional Key Words and Phrases: Cognitive biases detection, cognitive-aware systems, eye tracking, gaze features

1 INTRODUCTION

Digital technology has enabled us to access virtually any information at any point in time. And while there is a sheer unlimited amount of information accessible, humans have only limited processing resources and abilities [25]. Meanwhile, massive amounts of data are collected and made available to people, making it more cognitively challenging to process them and make informed decisions based on this data [21]. Consequently, people use heuristics, so called cognitive shortcuts, to reduce the complexity of information-based decision making. In short, heuristics describe the use of previously developed attitudes or perspectives to make the decision making process easier, rather than fully (re)evaluating a situation systematically and rationally.

In 1974, Tversky and Kahneman proposed the concept of cognitive bias and identified various heuristics that could cause bias [67]. Cognitive bias is a systematic pattern of deviation from norm or rationality in judgment [25]. To be specific, when making decisions or judgments, individuals who exhibit cognitive bias tend to follow their own beliefs or preferences rather than objective information [25]. One of the most prominent examples of cognitive bias is the confirmation bias, where people only search for or evaluate information that supports their existing perspectives or expectations [40, 52]. For example, confirmation bias might exist in criminal evidence collection, as the investigator only seeks or overly emphasizes the data that supports the guilty claims because he/she believes so. Conversely, the inclination of avoiding resolving conflict and reducing dissonance is known as cognitive dissonance [19].

While sometimes heuristics like confirmation bias are desirable as they enable fast and effective decision [67], they also make people susceptible to the spread of misinformation and malicious guidance. Polarization and user

Authors' addresses: Xiuge Chen, xiugec@student.unimelb.edu.au, The University of Melbourne, Melbourne, Victoria, Australia; Benjamin Tag, benjamin.tag@unimelb.edu.au, The University of Melbourne, Melbourne, Victoria, Australia; Catherine E. Davey, catherine.davey@unimelb.edu.au, The University of Melbourne, Melbourne, Victoria, Australia; Tilman Dingler, tilman.dingler@unimelb.edu.au, The University of Melbourne, Melbourne, Victoria, Australia.

segregation online have been shown closely related to misinformation because of the lack of supervision on the contemporary internet [13, 77, 78]. Recent studies have demonstrated that when facing conspiracy or scientific topics, Facebook users tend to view information that supports their beliefs, even when it is deliberately false [76, 78] . Meanwhile, dissenting information will either be ignored or result in an unexpected backfire effect, causing an increase in their activities related to their beliefs [76]. Our current internet and technologies have facilitated this process even further. The ease of discard information in digital world made people more exposed to confirmation bias and cognitive dissonance [6]. Furthermore, the advance of personalized algorithms and systems have been accused for indirectly contributing in reinforcing people's inherent biases because of their effort in keeping users engaging [5]. This, as a result, helped the creation of echo chambers [56] and the spread of disinformation [72]. In most cases, confirmation bias is inherent and subconscious [67], people are unaware that they are making irrational decisions or judgments. Therefore, to mitigate the effect of misinformation and depolarize the society, it is crucial to be able to inform people about the presence of confirmation bias and cognitive dissonance.

In this work, we designed an experiment that could successfully induce confirmation bias and cognitive dissonance, and conducted an in-lab experiment collecting eye movement data from 33 participants. We extracted gaze features from the raw gaze data and trained classifiers to differentiate information sources that either aligned with or contradicted against participants' prior beliefs, based on their self-assessed attitudes. Our models achieved up to 76% accuracy for text stimuli under a generic evaluation method, which improved previous results in terms of both performance and generalizability. Combining with our analysis of the critical gaze differences, this study facilitate the progress of building cognitive-aware systems, i.e., systems that can take users' cognitive state into consideration and act correspondingly to improve user experiences. With our biases detection techniques, systems can work interactively with users on mitigating biases and misinformation spread, and ultimately, develop a more informed and depolarized society.

2 RELATED WORK

Researchers have put much effort into detecting the presence of confirmation bias and cognitive dissonance. Broadly speaking, these studies can be categorized into two types, perception-driven approach and behavior-driven approach. The former method focuses on how people perceive the information sources, while the latter utilizes the unconscious behavioral data including interactions and biophysical responses.

2.1 Perception-driven Approach

Perception-driven approaches are developed based on the observation that people incline to trust more on attitude-consistent information [14, 40, 52], i.e., information which aligns with their prior beliefs. Example metrics include information importance rating (by participants) and the degree that selected information agrees or contradicts with one's belief [55], etc. These metrics have been used to indicate the presence of cognitive biases in many studies [20, 30, 35, 36, 39, 49, 68, 75]. Although perception data is easy to collect, it can not be used to reflect bias in real time [53] as they usually involve post-study data collection (e.g., post-questionnaire [20]) and analysis. Moreover, these measurements can be hindered by subjective factors like participants' attitude or awareness [26].

2.2 Behavior-driven Approach

Since the presence of cognitive biases is subconscious, people who exhibit bias might behave differently or have different behavioral characteristics [52, 55]. Processing information that supports ones pre-existing opinion is more enjoyable [48] and requires less cognitively effort [32]. This, as a result, has been used to indicate the presence of biases during a variety of tasks.

The simplest but also widely used behavioral feature is task completion time [3, 66] and reaction time [15, 22]. As a result of confirmation bias, people tend to read and spend longer time on sources that align with their pre-existing beliefs [3]. Meanwhile, they may prefer seeking out attitude-consistent information over attitude-contradicting information, which is also known as selective information search [3, 37]. Moreover, in visual analytics where people interact with visualized information through digital inputs (i.e., keyboard and mouse), interaction data can also be used to identify biases. These interactions are central to the visual analytic systems [17] and can be used to reflect one's decision process [57]. Cho et al. [11] found that there is a significant difference in the interaction pattern between biased and unbiased participants. Meanwhile, Nussbaumer et al. [53] and Wall et al. [71] have proposed theoretical frameworks and metrics that use only interaction data to detect cognitive bias in visual analytics. In general, both data interaction time and frequency could play an essential role in identifying bias. These metrics have been further refined and implemented in various settings to estimate confirmation bias [50] and other cognitive biases, such as anchoring bias [18, 69, 70] and selection bias [23]. However, these methods are constrained by the existence of computer-based interaction data – there could be no such data at all in many other scenarios, such as online news and social media. Therefore, to battle misinformation, other indicators need to be identified.

Biophysical approaches have several advantages over behavioral approaches, not only because they do not suffer from subjectivity and delayed reflection but also because they are unconscious features that would appear in almost every context. Typically, people capture information via vision when there are no such specially designed visualized systems. Since the eye-tracking technique provides measurements for memory loading [44] and attention levels [31], it becomes a natural candidate for bias estimation. To illustrate, similar to the task completion time, viewing duration could reflect the efforts that participants put in processing information. Several researchers have investigated the eye patterns under the context of Information Retrieval (IR). In 2015, Schneider et al. [60] suggested that gaze data like fixation points and focus time could be used in identifying three types of bias in online review forms, but no actual experiments have been performed. Similarly, Shokouhi et al. [61] and Bhattacharya and Gwizdka [7] pointed out that the amount of information acquired is associated with the duration spent and different eye movement patterns. For instance, when obtaining more knowledge, sequential fixations and regressions tend to be longer. Although these studies did not directly assess the correlation between gaze movements and cognitive bias, their results suggest that eye-tracking is a promising tool for estimating information loading. As many biases are related to different degrees of information processing, it is not surprising to start exploring the possibility of using gaze patterns to estimate cognitive bias.

There are only a few studies that have studied the relationship between eye movements and confirmation bias. Several scientists studied users' attention and confirmation bias in online political poster advertisements [46, 59, 74]. During the studies, participants were asked to choose advertisements that either agreed or disagreed with their political party preferences. Results demonstrated both the existence of confirmation bias and its correlation with the staring time. Participants tend to select and spend more time focusing on the advertisements that align with their political status. However, Sülflow et al. [63] failed to observe such a relationship. In their study, participants were exposed to Facebook news feeds that either support or against the German refugee policies. Although users still selected the news confirming their status, they spent an equal amount of time on both types. It is worth noticing that this result does not invalidate the previous result [46, 59, 74], as the opposite conclusions are drawn from different study designs. To explain, the first three studies [46, 59, 74] were conducted with party election topics in the form of advertisements. In contrast, the last study [63] was performed under government policy news feed (on Facebook) and only undergraduate students were recruited as their participants. Hence, further explorations need to be done to determine the relationship between confirmation bias and viewing time in different scenarios.

Instead of confirmation bias, other research examined the eye patterns in other cognitive biases that are also accompanied by different attention levels. For example, several scholars [1, 2, 16, 42, 45] studied attentional bias,

a bias caused by special attention to partial information, in depression and dysphoria. Compared to healthy participants, depressed or dysphoric participants demonstrated attentional biases towards depression-related stimuli. They tended to spend more time viewing the negative (depression-related) images and spend less time watching the positive images. Similarly, Harris [24] studied the relationship between gaze time in Areas of Interest (AOI) and another bias, the bandwagon effect. The bandwagon effect happens when people overvalue unrelated information in their decisions. When participants exhibit the bandwagon effect, they spend significantly more time in areas that contain unrelated information. Meanwhile, they spend much less time concentrated on areas that include the actual task. In conclusion, these studies suggest that the fixation duration could be used to reflect the amount of cognitive efforts. Because the presence of confirmation bias is also related to the increasing amount of cognitive efforts [67], these results also suggest that eye-tracking techniques could be used to estimate confirmation bias.

Despite the rich history of researching eye movements in cognitive biases, to the best of our knowledge, no studies have been done in utilizing gaze data to quantify or classify the presence of cognitive biases. Recently, Villarreal et al. [68] did a pilot study to explore the possibility of using a different biophysical signal, EEG, to predict the presence of confirmation bias. They trained a machine learning model where the input is EEG signals and the ground truth is the behavior-based result. Although their model had plausible performances on some participants, their results were obtained by within participant evaluation with a small number of participants. Hence, further exploration needs to be done to assess the generalizability of their results.

Therefore, drawing on the research above and the gaps they are not able to address, we form our research questions as below:

(1) Can we design systematical studies to observe the presence of confirmation bias and cognitive dissonance?
(2) Can we use eye movements data to detect the presence of confirmation bias and cognitive dissonance?

3 DATA COLLECTION

We described our study design and procedure in this section. Differ from similar studies [46, 49, 59], we selected our stimuli from a wider range of topics. In short, we carefully chose texts stimuli for four noted topics with commonly polarized groups. We ensured that all stimuli delivered a clear message, and were displayed in a consistent format. Then, we presented our experiment setup and procedure, as well as an overall summary of participants' demographic attributes.

3.1 Stimuli Selection

Our topic selection was based on several criteria: They should be well-known to most participants in Australia; Participants were likely to have a strong opinion on them; and they covered a wide range of areas, including science, politics, societies, and cultures. As a result, we selected four topics to design our stimuli, which were feminism versus anti-feminism, progressivism versus conservatism, multiculturalism versus Australian traditional culture, and the main cause of climate change (i.e., man made versus natural cycle).

All four topics were well-known to the public with increasingly polarized standpoints. For instance, political attitudes about progressivism and conservatism in western society have become more and more polarized since the 1970s [8, 10, 47, 73], which is reinforced by personalized algorithmic technologies to create echo chambers and filter bubbles [51]. Similarly, we chose feminism because of the rising responses from some man's groups [58] and third-wave feminists [64]. Moreover, the conflict between multiculturalism and Australia's own culture has existed since the "White Australia Policy" [12], an immigration restriction policy that targets against non-British citizens, making it a suitable topic for Australia-based studies. Lastly, despite the fact the majority supports that humans should be primarily responsible for the current climate change [33], strong rejections still exist in many

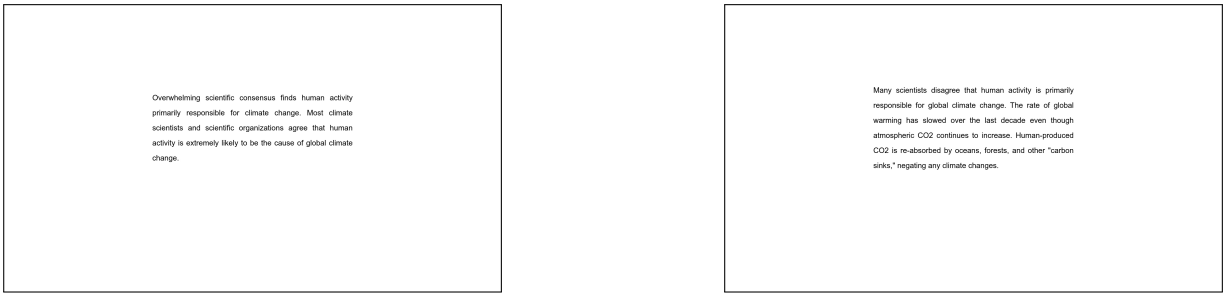
groups like conservative and free-market ideology [43]. The large discrepancy between the science community and climate change denial society has made it appropriate for bias-related studies.

As shown in Table 1, for each pair of topics except climate change, we chose four specific subtopics to design stimuli. Four researchers were involved in the subtopic and stimuli selection, and evaluated the stimuli independently.

Table 1. Subtopic Selection for each Topic

Topic	**Subtopic 1**	**Subtopic 2**	**Subtopic 3**	**Subtopic 4**
Feminism	Abortion rights and reproductive freedom	Women's rights	Workplace equality	Well-known Slogan
Anti-feminism	Anti-abortion	Men's rights	Workplace inequality	Well-known Slogan
Progressivism	Women in military	Gay marriage	Mardigras parade	Australia Labor party
Conservatism	Anzac badge	Traditional family	Sheep farm	Australia Liberal party
Man-made	Posters and protests			
Natural cycle	Posters and protests			
Multiculturalism	Mixed sports team	Immigrant protest	Multicultural festivals	International students
Australian Culture	Cricket	Scott Morrison	Australian BBQ	HMS endeavor

We select one text for each subtopic as our stimuli. And we made our best effort to ensure the consistency across our stimuli so that participants' should present similar eye movements across all stimuli if they hold a neutral position. Specifically, all texts were 50 words and contained no complex words or sentences. As shown in Figure 1, the same font (Arial 30 px), line spacing (double), alignment (justified and centered), and column width (800 px) were chosen for each text.



(a) Man Made (b) Natural Cycle

Fig. 1. Examples of Text Stimuli and Their Display.

3.2 Experimental Setup

We recorded participants' eye movements through the Tobii Pro X3-120 eye tracker [1]. And we used the associated software, Tobii Pro Studio, to help design the study and preprocess the raw gaze data. The eye tracker had a sampling rate of 120 Hz. The output of Tobii Pro Studio included both the raw gaze data (e.g., (x, y) coordinates on the screen, and left/right pupil diameters), and the preprocessed fixation and saccade data. We mounted the eye-tracker at the bottom of a 24-inch monitor.

3.3 Participants

Thirty-three participants (19 women, 14 men) were recruited for this in-lab study, all of whom were students and staffs from our university. All participants were native or bilingual English speakers. Among the 31 participants who had specified their ages, two were under 20, sixteen were between 20 and 30, three were between 40 and 40, seven were between 40 and 50, and the rest three were above 50. Most participants had an education level of postgraduate degree ($N = 15$), followed by bachelor degree ($N = 10$), while the remaining were graduate certificate or diploma ($N = 3$), certificate III or IV ($N = 1$), and year 12 ($N = 2$).

3.4 Procedure

To begin with, participants were informed about the purpose and procedure of this study. If would like to proceed, participants were asked to sign a consent form and fill a prestudy survey regarding their demographic attributes and some general questions about each topic (e.g., familiarity, opinions). Next, we seated them in a comfortable position and asked them to adjust the seat such that their heads were centered and approximately 60 − 65 cm away from the Desktop screen. During the stimuli viewing session, a trial topic on flowers versus insects was displayed first, and followed by the four topics of focus in counterbalanced order. Within each topic, stimuli were displayed in random order without intervals in between. Before viewing each topic, the eye tracker was calibrated through a standard built-in 9-point calibration procedure. While viewing the stimuli, participants were instructed to try not to move their head around, and click the keyboard using the other hand to proceed to the next stimuli. Lastly, participants were asked to finish a post-study survey about their experiences in this study. It took around 45 − 60 minutes to finish the study and upon completion, participants were compensated with a 20 dollar coupon.

4 METHOD

In this section, we describe our approach for classifying attitude aligning versus contradicting stimuli in detail. We used the preprocessed fixation data by the Tobii Pro Studio and treated the transitions between fixations as saccades. Unlike other eye studies, we did not partition data into equivalent-length time windows, since task completion time has been shown to be correlated with information alignment [66]. Instead, we simply take the data from one stimulus as one instance. We then extract 25 low-level and 50 mid-level [62] features for each data instance as the model input. Afterwards, we developed and tested the performance of several classification models with repeated nested cross-validation [9]. Within the inner cross-validation on the training set, we performed supervised filter-based feature selection and hyper-parameter tuning for each model, whose results were used in the later outer cross-validation.

4.1 Feature Extraction and Selection

4.1.1 Feature Extraction. Feature extraction is required in this study because the raw and preprocessed data are not sensible inputs to the model and vary in dimensions. Classic feature extraction techniques categories gaze features into three levels, low-level, mid-level, and high-level [62]. Low-level features are most widely used

[1] https://www.tobiipro.com/product-listing/tobii-pro-x3-120/

in eye-tracking studies [27, 34], which mainly contain statistical descriptions of raw and preprocessed data, such as the count of fixations, mean and variances of saccades, etc. Mid-level features, on the other hand, focus on the small local movement trends among consecutive fixations [62]. By attributing these movements into different types (e.g., *line reading*, scan, and comparison), one can generally obtain a better view about the reading behaviors. High-level features extract information from the area of interest (AOI), like the duration spent in AOI and the count of in/out saccades. Since identifying AOI requires domain knowledge and is specific to individual stimuli, we chose not to include high-level features for generalizability. As detailed in Table 2 and Table 3, we extracted 25 low-level and 50 mid-level features.

In low-level features, we calculated the *rate* feature by dividing the sum of fixation duration by the total amount of time spent on this task (i.e., stimulus). Similarly, we defined the *percentage* to be the ratio between fixation count and task duration. We derived the fixation *slope* by fitting a regression line to the fixation plot, and the 75% *dispersion area* by finding the smallest area containing at least 75% fixations. For saccade features, we considered the directions of each saccade and the relationship with the previous saccade. For instance, *up* saccades were those pointed upwards and within 22.5 degree from the vertical line. And two consecutive were counted as one *neighbor* if they had adjacent directions, respectively. Moreover, we categorised fixations and saccades into *small*, *medium*, and *long* based on their durations and lengths. For fixations, we used duration thresholds of 200 and 500 ms, and for saccades, we used length thresholds of 100 and 400 px, so that long saccades were more than half of the text column width [2]. Our mid-level feature extraction was identical with Srivastava et al. [62].

Table 2. Selected Low-level Features

Category	Features	Measurement
Fixation Duration	mean, std, var, min, max	value
	rate, percentage, slope, dispersion area (75%)	value
	all, short, medium, long	count
Saccade Length	mean, std, var	value
	follow, neighbor, opposite, opposite-neighbor	count
	right, left, up, down, up-right, down-right, up-left, down-left	short, medium, long count
Pupil Diameter	mean, std, var	left and right value

Table 3. Selected Mid-level Features. All features are measured by their count.

Category	Sub-category	Features
Shape Based	String	up, right, down, left
	Line	small, medium, long
	Comparison	left-right, right-left, up-down, down-up
	Scan	right-left, left-right, up-down, down-up, right-up, up-right, left-up, up-left, right-down, down-right, left-down, down-left
Distance Based	-	regression, else-where

[2]Screen size: 1080 × 1920 px. Images were displayed in full-screen mode, and text column width was 800 px

4.1.2 Feature Selection. Feature selection is the process of selecting a subset rather than the full set of features as model input. We incorporated feature selection in this study to help reduce computation complexities and improve the generalizability of our trained models. We adapted supervised filter-based feature selection with different criteria outlined below, since it was efficient and decoupled with the models so that it is less likely to overfit. In filter-based selection, the importance of one feature can be quantified by a specific statistical correlation between it and the label. And we measured such correlation with four criteria: chi-squared statistics (χ^2), information gain (IG), joint mutual information (JMI), and double input symmetric relevance (DISR). The best subset of features is determined in a greedy forward manner. That is, we repeatedly added one most promising feature into our selection until there were no higher evaluation results for a while. For all criteria, we chose their evaluation metric to be the average accuracy of repeated cross-validation.

4.2 Classification

We defined our classification problem to be predicting whether a stimulus was supporting or contradicting one's own opinion. Hence, we labelled each data instance based on its alignment with participants' self-assessed opinions. For example, feministic participants had their feminism stimuli labelled as "Aligned" and anti-feminism stimuli labelled as "Contradicting". An instance was disregarded if the participant held a neutral position. Since our dataset is perfectly balanced, we evaluated the model performance via accuracy score, i.e., the proportion of correctly classified instances.

Baseline: Because of our balanced dataset, a trivial baseline would be an accuracy score of 0.5. In addition, previous studies demonstrated the strong correlation between task completion time and information source alignment [66]. Therefore, we also trained and tested our classifiers with a single feature, duration, and reported the results as a more realistic baseline.

In the training and testing process, we compared the performance of several classification models, including Linear and RBF Kernel Support Vector Machines (SVM), Logistic Regression (LR), Random Forest (RF), and XGBoost (XGB). Each model was evaluated through the leave-one-participant-out train and test approach. In each round, one participant's data was saved for testing, and data from the rest of the participants were used for training. We repeated this for each participant and reported the average accuracy. By doing so, we ensured that we maximized our training set without exposing the testing set into the training. Hence, our results could better imply the generalized performance on unseen participants. Moreover, model hyper-parameter tuning and feature selection could often improve model performance, but doing it against the testing dataset may lead to overfitted results. One commonly used solution for this is to divide the dataset into training, testing, and validation set, but this would significantly reduce the amount of available data for training. Due to the limited data in hand, we used nested cross-validation [9] to help address this problem, which simply nests the tuning steps within the training stage. In short, an outer cross-validation was performed on the whole dataset, which essentially is done by our leave-one-participant-out validation described above. For each validation round, an inner cross-validation was performed on the training set to perform tuning. We chose it to be a ten-fold cross-validation with ten repetitions, and used the average accuracy to select the best parameters, which later was used for testing. For Random Forest, we tuned the number of estimators, max depth, min number for sample split, and min number of sample leaf. For XGBoost, we tuned the number of estimators, max depth, min child weight, learning rate, subsample ratio, and subsampling of columns. For Logistic Regression and SVM, we tuned the solver/kernel, penalty types, and penalty strength.

Some participants may have stronger opinions than others, and attitudes on some topics may be more extreme than others. Participants with stronger opinions may demonstrate a more significant difference in responses when viewing different stimuli. Moreover, people's behavior might be influenced by the topic type, i.e., the overall

degree of polarization might be stronger in some topics. Consequently, we also evaluated model performance on only strong opinioned participants and on each topic, respectively.

Lastly, we took the best-performing model and analyzed it in more detail. More specifically, we looked at critical features in classification and compared them with previous known indicators. We also attempted to attribute reasons for misclassification, i.e., whether it is due to model inefficiency or the inherent variance among people.

5 RESULTS

We began this section with a summary of participants' self-assessed attitudes and how we categorized them into weakly and strongly opinioned. With a sufficient amount of polarized participants, we then demonstrated our promising results for topic-independent classification and topic-specific classification, respectively. Afterwards, we dug deeper into the best-performing model and obtained more insights on important features as well as potential explanations for misclassified instances.

5.1 Self-assessed Attitude

Participants were asked to assess their opinions on each topic and their strength on a scale of zero to five. We then categorize participants into neutrally, weakly, and strongly opinioned based on a threshold of 2. For example, in Figure 2, participants with 0 opinion strength were considered neutral, whose data for this topic would be removed during the evaluation stage. Participants with low opinion strength (i.e., had a non-zero score no higher than two) were categorized as weakly opinioned. At the same time, the rest of the participants were added to the strongly opinioned list. Table 4 summarized the count of participants by opinions and by topics. Although we tried to recruit participants from diverse backgrounds, most participants had strong opinions towards one side.

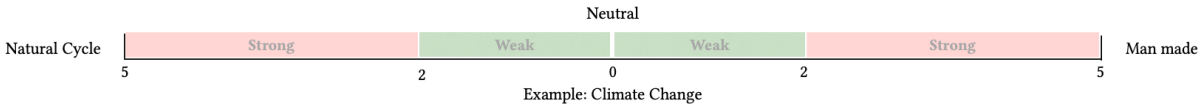


Fig. 2. Example of inclination score categorization.

Table 4. Self-assessed Attitude for Each Topic. See Table 1 for each topic's full name.

Opinion	Climate Change		Gender		Politics		Culture	
	Man-made	Natural	Fem.	Anti-fem.	Pro.	Con.	Mul.	Aus.
Neutral	2		1		3		2	
Weak	6	1	8	2	11	1	5	0
Strong	24	0	20	2	16	2	25	1

5.2 Topic-independent Classification

As indicated in Section 4, we evaluated our topic-independent classifiers by aggregating data from all participants from all topics and removing neutrally positioned data. Meanwhile, we also tested with removing both weakly and neutrally positioned data. We used the majority guesser (0.5 accuracy) and classifiers with a single feature, duration, as our baselines. We reported the best classifier and accuracy score for each condition in Table 5.

Table 5. Topic-independent Classification Results.

Opinion	All		Strong-only	
Best	**Model**	**Accuracy**	**Model**	**Accuracy**
Duration	Linear SVM	0.5323	Linear SVM	0.5938
Gaze + Duration	RF	0.6461	RF	0.6839

Table 6. Topic-specific Classification Results.

Topic	Climate Change		Gender		Politics		Culture	
Opinion	**All**	**Strong-only**	**All**	**Strong-only**	**All**	**Strong-only**	**All**	**Strong-only**
Best Model	RF	XGB	XGB	RBF SVM	RF	RF	LR	RF
Best Accuracy	0.7150	0.7614	0.6385	0.6758	0.6596	0.7344	0.6671	0.7045

Comparing the two baseline approaches, we found that task completion time had limited power in classifying aligned versus contradicting sources. And it was more useful in texts than in images. With duration feature only, we could achieve an accuracy score of 53.23%, which could be boosted to 59.38% when only including strongly opinioned participants. The usefulness of duration could also be observed when analyzing the statistical correlation between duration and task labels. That is, duration had a statistically significant relationship with our stimuli labels (paired t-test $t(992) = 5.42, p < 0.001$).

As shown in Table 5, when combining gaze data with task completion duration, our classifiers significantly outperformed the two baselines. Random Forest achieved the best performance in all scenarios. It obtained an accuracy score of 64.61% on non-neutral participants and 68.39% on strongly opinioned participants. By including only strongly polarized participants, the accuracy increased in all cases. This aligned with one of our hypotheses that strong polarization may make participants rely more on taking cognitive shortcuts, which was then reflected in the more obvious biophysical responses.

5.3 Topic-specific Classification

In addition to the generic model described in Section 5.2, we also experimented with topic-specific classifiers. Instead of concatenating all data together, we evaluated our models on each topic separately using the same methodology. Similarly, we reported the model with the best accuracy in Table 6.

As indicated in Table 6, we received a significant increase in accuracy score after separating the topics. In all except one case, our best classifier could identify over 70% of instances correctly, which then can be boosted to up to 76.14% when including strong polarized participants only. Random Forest remained the best performing model in general, but it had also been beaten by other models in some scenarios. All models performed significantly better in politics and science (i.e., climate change) topics than the other two topics, indicating strong signals for confirmation bias and cognitive dissonance. One possible explanation is that these two topics were more polarized in our participant pool, but it is not reflected in the self-assessed results.

5.4 Important Features and Misclassifications

In this section, we analyzed Random Forest's classification behavior in detail, since it generally outperformed other models in both topic-independent and topic-specific studies. Random Forest rates feature importance by

Table 7. Top Ten Important Features by Random Forest.

Rank	**1**	**2**	**3**	**4**	**5**
Feature	Left long sacc.	Duration	Fix. slope	Medium right sacc.	Medium fix.
Rank	**6**	**7**	**8**	**9**	**10**
Feature	Opposite sacc.	Sacc. mean	Sacc. var	Sacc. std	Fix. number

Gini importance score, which is the number of a feature being used in node splitting weighted by the number of samples being split. Hence, to study how Random Forest made predictions, we fitted all data into the Random Forest model and reported its top ten features in Table 7.

As demonstrated in Table 7, Random Forest used mostly saccade-related features. Consistent with previous studies [46, 59, 74], participants tended to spend longer time on aligned sources, resulting in more fixations and saccades. More importantly, the classification relied heavily on several reading-related features, such as fixation slope, long left saccades, and medium right saccades. There were more regression patterns and repeated readings when reading texts with an aligned view. As indicated in prior information processing studies [7, 61], these observations revealed that participants spent more effort in processing information with aligned views. Hence, the presence of confirmation bias and cognitive dissonance during our experiments could be validated.

However, the observation above did not fit all participants. Instead, some participants tended to spend more time on contradicting sources, resulting in more fixations and saccades, as well as re-reading alike patterns in stimuli with opposite views. Similarly, some participants spent a roughly equal amount of time and behaved quite the same in all stimuli. Consequently, their data contributed to many of the misclassified instances. Examples of such participants and participants with expected behaviors were shown in Figure 3. Participant 14 (top row) behaved as expected in culture stimulus, while participant 26 (bottom row) demonstrated opposite behavior in climate change stimulus. The unexpected behaviors might be caused by inaccurate self-assessed attitudes, or the overrode of information utility. Information utility captures participants' evaluation of information usefulness for future decisions. When participants perceive the information as useful, they will be likely to engage with the information regardless of its status [38]. Hence, further studies were required to investigate the misclassified instances in detail.

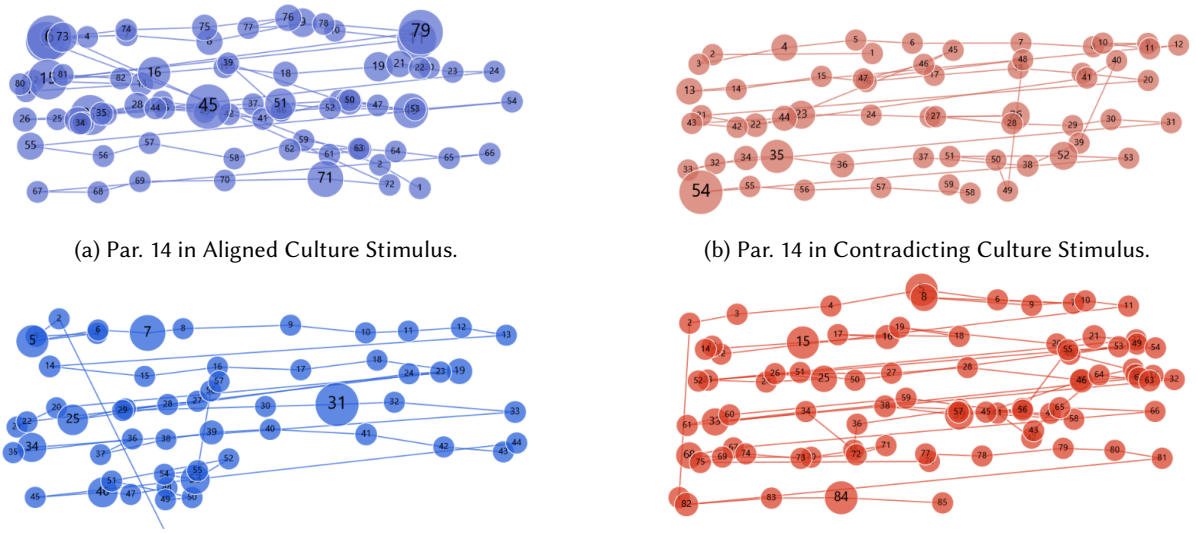
6 DISCUSSION

To mitigate the effect of cognitive bias during information consumption and decision-making, we need to be able to quantify its presence effectively and non-intrusively. In this study, we developed a study design that can systematically induce confirmation bias and cognitive dissonance. More importantly, we demonstrated the correlation between such biases and eye movements, and built a generic high-performing classifier that can differentiate aligned and contradicting sources.

6.1 Eye Movements and Cognitive Biases

By analyzing the eye movement map and the important features rated by the Random Forest classifier, we were able to link the connections between gaze and cognitive biases. More specifically, we showed that participants' gaze moved differently when viewing aligned and contradicting information sources.

In general, participants tended to spend longer time on the stimulus that agreed with their prior beliefs. Consequently, there were more fixations and saccades in the gaze pattern. Moreover, participants had more short saccades and more regressions in reading aligned stimuli, indicating a heavier information processing stage. While



(a) Par. 14 in Aligned Culture Stimulus.

(b) Par. 14 in Contradicting Culture Stimulus.

(c) Par. 26 in Aligned Climate Change Stimulus.

(d) Par. 26 in Contradicting Climate Change Stimulus.

Fig. 3. Example Gaze Movements for Text Stimulus.

in contradicting sources, most of the saccades were long saccades, and there were barely any revisions. These reading pattern differences were less obvious in images than in texts, since they had more varied presentations causing participants to process them differently. Also, some participants demonstrated opposite behaviors. They spent significantly longer time as well as more cognitive loads in contradicting sources. This might be due to the inaccuracy of self-assessed answers, or participants' intentions in treating both views equally (i.e., override of information ultility [38]), or their unfamiliarity with contradicting sources.

6.2 Cognitive Biases Detection

In this study, we showed a promising way to detect cognitive biases using eye movement data. The differences in processing aligned and contradicting sources allowed us to identify participants' own beliefs, hence making it feasible to detect confirmation bias and cognitive dissonance.

In the most general setting where we included data from all topics and all non-neutral participants, we achieved an accuracy score of 65% And by including heavily polarized participants only, the accuracy could be boosted up to 68% Across all the classifiers, Random Forest generally beat others. And all of the classifiers significantly outperformed our choices of baseline, the majority guesser (accuracy 50%) and the models trained with single feature duration (53%). Therefore, eye movement data turned out to be much more helpful than the duration, even though many gaze features were highly correlated with duration itself. This was because, in addition to the time spent, eye movements data also revealed how participants process the information, i.e., the actual cognitive load.

Moreover, when trained and tested classifiers for each topic separately, i.e., topic-specific classifications, our models can achieve better performance. The best performing models could distinguish over 65% of the instances correctly for all topics except *Gender* (64%). The model accuracy could be further increased to above 70% when including strongly opinioned participants only (except for *Gender* which had 68%). For the most accurate topics, *Climate Change* and *Politics*, our models can achieved an accuracy score of up to 76% and 73%, respectively. This

indicated that participants presented cognitive biases more easily while consuming information under these two topics.

After looking into the misclassified instances in detail, we argued some of the misclassifications were unavoidable due to the inherent variance within participants and our choices of ground truth. As mentioned in Section 6.1, some participants exhibit opposite behavior. That is, they spent more effort on sources contradicting their prior beliefs, causing the inaccurate classifications of our models. One possible explanation was the subjectiveness of our choice of ground truth, which may not reflect participants' true attitudes. Moreover, previous studies [38] indicated the possibility of information utility override, where participants process information regardless of its attitude because they perceive it as useful. Further studies need to be done to investigate them in detail.

Our results outperformed previous studies in terms of both accuracy and generalizability. Hence it could be applied to general settings with new topics and new participants without much performance degradation. While previous studies generally used one topic with one presentation, and within participants evaluation [46, 49, 59, 68]. We chose the leave-one-participant-out evaluation approach combined with nested cross-validation for parameter tuning, so that the testing data was always an unseen participant and was not exposed to the tuning process. Additionally, we chose a variety of topics and stimuli, and trained our generic model with all data as detailed above.

6.3 Limitations

To begin with, we collected our data from an in-lab study where participants performed studies in controlled settings. For instance, participants were instructed not to move their heads, and they were aware that they were observed by the experimenter. Moreover, participants were informed of the study purpose and procedure prior to the commencement of the study. Hence, their might behave differently, making the collected data does not reflect the real-world data perfectly. However, we argued that this stationary setup was necessary to investigate the relationships between gaze patterns and cognitive biases. And our results should be viewed as the foundations for future studies.

Moreover, we used self-assessed attitudes as our ground truth, which was not fully reliable. To evaluate the models more accurately, better ground truth was required. This can be done with an extra validation stage with expert input or other well-studied measurements, such as the affinity score [30, 35, 36, 39, 49, 75]. In addition, even though we had tried to recruit participants from various backgrounds, our participants' profile was still highly biased. We had much more participants in favor of one side than the other side. Also, some of our images stimuli were ambiguous to participants, so that they might perceive them incorrectly. For instance, to represent multiculturalism, we selected an image with children of different colors playing Australian football together. But some participants found it demonstrated more Australian culture as Australian football was one of the most popular sports in Australia.

During the feature extraction, our choices of thresholds for low-level and mid-level features were arbitrary. We simply chose them based on text width and screen width. By choosing the features more carefully, or even designing new features specifically for our stimulus, we could potentially improve the model performance. But this often came with the price of generalizability. Hence, this strategy should be used with cautiousness to prevent overfitting.

6.4 Future Directions

The future directions of this work can be summarized into three categories: Improving model performances, generalizing results, as well as applying our findings to similar studies.

There are many different ways we can experiment to improve classification efficiencies. With the gaze data, one can attempt to perform deeper feature engineering to extract more useful information from the gaze data.

This includes new features in all three levels (i.e., low, mid, and high) as well as adjustment to current features (i.e., thresholds in determining long saccades). Alternatively, we can include other biophysical data, such as using EEG, fNIR, EDA, heart rate, and temperature data. EEG signals have been used in detecting cognitive biases [68], and forehead temperature is shown to be correlated with EEG activities [28, 29]. Similarly, some works used EDA in studying biases [4, 48, 49]. In addition, several studies have suggested how heart rate variability is related to cognitive processes [41, 65] and hostility attribution bias [54]. Hence, with a more careful study design, it is expected that better performance can be obtained with more types of data.

Also, we can explore how to generalize our results to the wild. This includes adding stimulus from other topics, getting inputs from experts on stimulus design, recruiting participants from diverse backgrounds, and conducting experiments in less controlled settings.

Another direction would be applying our results to similar future studies. With an efficient classifier, one can systematically quantify the degree of cognitive biases during information consumption, which is very useful in many studies. For example, in bias mitigation studies, we can quantify the effect of various mitigation methods and design better solutions to support rational decision-making.

7 CONCLUSIONS

While enjoying the benefit of cognitive shortcuts, heuristics, we must address their side effects – partial information evaluation and irrational decision making. In this study, we investigated two common cognitive biases, confirmation bias and cognitive dissonance, in digital information consumption. We designed eye-tracking experiments that could successfully induce the presence of such biases, and conducted an in-lab user study with 33 participants. Our analyses on the eye movements revealed a close relationship between gaze patterns and information source alignment. Moreover, we demonstrated that, with classifiers trained on the extracted gaze features, we could effectively identify whether participants were viewing aligned or contradicted sources (compared to their prior beliefs). Lastly, we discussed the key features in helping differentiate information sources, and potential explanations for misclassified instances. Our work built an important foundation for cognitive biases detection through implicit biophysical responses, which paved the way for future related studies on understanding various biases and mitigating their effects on our life.

REFERENCES

[1] Kimberly Albert, Violet Gau, Warren D Taylor, and Paul A Newhouse. 2017. Attention bias in older women with remitted depression is associated with enhanced amygdala activity and functional connectivity. *Journal of affective disorders* 210 (2017), 49–56.

[2] Jody E Arndt, Kristin R Newman, and Christopher R Sears. 2014. An eye tracking study of the time course of attention to positive and negative images in dysphoric and non-dysphoric individuals. *Journal of Experimental Psychopathology* 5, 4 (2014), 399–413.

[3] Karl Ask and Pär Anders Granhag. 2005. Motivational sources of confirmation bias in criminal investigations: The need for cognitive closure. *Journal of investigative psychology and offender profiling* 2, 1 (2005), 43–63.

[4] Julia Avram, Felicia Rodica Balteş, Mircea Miclea, and Andrei C Miu. 2010. Frontal EEG activation asymmetry reflects cognitive biases in anxiety: evidence from an emotional face Stroop task. *Applied psychophysiology and biofeedback* 35, 4 (2010), 285–292.

[5] Alexander Benlian. 2015. Web personalization cues and their differential effects on user assessments of website value. *Journal of management information systems* 32, 1 (2015), 225–260.

[6] W Lance Bennett and Shanto Iyengar. 2008. A new era of minimal effects? The changing foundations of political communication. *Journal of communication* 58, 4 (2008), 707–731.

[7] Nilavra Bhattacharya and Jacek Gwizdka. 2018. Relating eye-tracking measures with changes in knowledge on search tasks. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*. 1–5.

[8] Filipe R Campante and Daniel A Hojman. 2013. Media and polarization: Evidence from the introduction of broadcast TV in the United States. *Journal of Public Economics* 100 (2013), 79–92.

[9] Gavin C Cawley and Nicola LC Talbot. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *The Journal of Machine Learning Research* 11 (2010), 2079–2107.

[10] Pew Reseach Center. 2014. Political polarization in the american public. *Retrieved September* 2 (2014), 2019.

[11] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 116–126.

[12] Chad Cooper. 2012. The immigration debate in Australia: from federation to World War One. (2012).

[13] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the National Academy of Sciences* 113, 3 (2016), 554–559.

[14] Patricia G Devine, Edward R Hirt, and Elizabeth M Gehrke. 1990. Diagnostic and confirmation strategies in trait hypothesis testing. *Journal of Personality and Social Psychology* 58, 6 (1990), 952.

[15] Catherine Donaldson, Dominic Lam, and Andrew Mathews. 2007. Rumination and attention in major depression. *Behaviour research and therapy* 45, 11 (2007), 2664–2678.

[16] Almudena Duque and Carmelo Vázquez. 2015. Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of behavior therapy and experimental psychiatry* 46 (2015), 107–114.

[17] Alex Endert, William Ribarsky, Cagatay Turkay, BL William Wong, Ian Nabney, I Díaz Blanco, and Fabrice Rossi. 2017. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*, Vol. 36. Wiley Online Library, 458–486.

[18] Mi Feng, Evan Peck, and Lane Harrison. 2018. Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 501–511.

[19] Leon Festinger. 1957. *A theory of cognitive dissonance*. Vol. 2. Stanford university press.

[20] Peter Fischer, Eva Jonas, Dieter Frey, and Stefan Schulz-Hardt. 2005. Selective exposure to information: The impact of information limits. *European Journal of social psychology* 35, 4 (2005), 469–492.

[21] John R Gersh and Nathan Bos. 2014. Cognitive and organizational challenges of Big Data in Cyber Defense. In *Proceedings of the 2014 Workshop on Human Centered Big Data Research*. 4–8.

[22] Ian H Gotlib and Douglas B Cane. 1987. Construct accessibility and clinical depression: A longitudinal investigation. *Journal of abnormal psychology* 96, 3 (1987), 199.

[23] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 85–95.

[24] Christopher G Harris. 2019. Detecting cognitive bias in a relevance assessment task using an eye tracker. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. 1–5.

[25] Martie G Haselton, Daniel Nettle, and Damian R Murray. 2015. The evolution of cognitive bias. *The handbook of evolutionary psychology* (2015), 1–20.

[26] William L. Hays. 1983. Experimental Design: Procedures for the Behavioral Sciences. 2nd ed. *Psyccritiques* 28 (1983).

[27] Shoya Ishimaru, Kensuke Hoshika, Kai Kunze, Koichi Kise, and Andreas Dengel. 2017. Towards reading trackers in the wild: detecting reading activities by EOG glasses and deep neural networks. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 704–711.

[28] SD Jenkins and RDH Brown. 2014. A correlational analysis of human cognitive activity using Infrared Thermography of the supraorbital region, frontal EEG and self-report of core affective state. In *Comunicación presentada en la 12ª Conferencia Internacional de Termografia*

de infrarrojo cuantitativa, Burdeos, Francia.
[29] Sean Jenkins, Raymond Brown, and Neil Rutterford. 2009. Comparing thermographic, EEG, and subjective measures of affective experience during simulated product interactions. *International journal of Design* 3, 2 (2009).
[30] Benjamin K Johnson, Rachel L Neo, Marieke EM Heijnen, Lotte Smits, and Caitrina van Veen. 2020. Issues, involvement, and influence: Effects of selective exposure and sharing on polarization and participation. *Computers in Human Behavior* 104 (2020), 106155.
[31] Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: From eye fixations to comprehension. *Psychological review* 87, 4 (1980), 329.
[32] Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review* 93, 5 (2003), 1449–1475.
[33] Natasha Kassam. 2019. Lowy Institute Poll 2019. (2019).
[34] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131.
[35] Antino Kim and Alan Dennis. 2018. Says who?: How news presentation format influences perceived believability and the engagement level of social media users. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.
[36] Antino Kim, Patricia L Moravec, and Alan R Dennis. 2019. Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems* 36, 3 (2019), 931–968.
[37] Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation* 32 (1995), 385–418.
[38] Silvia Knobloch-Westerwick and Steven B Kleinman. 2012. Preelection selective exposure: Confirmation bias versus informational utility. *Communication research* 39, 2 (2012), 170–193.
[39] Silvia Knobloch-Westerwick, Ling Liu, Airo Hino, Axel Westerwick, and Benjamin K Johnson. 2019. Context impacts on confirmation bias: Evidence from the 2017 Japanese snap election compared with American and German findings. *Human Communication Research* 45, 4 (2019), 427–449.
[40] Asher Koriat, Sarah Lichtenstein, and Baruch Fischhoff. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human learning and memory* 6, 2 (1980), 107.
[41] Sylvia D Kreibig. 2010. Autonomic nervous system activity in emotion: A review. *Biological psychology* 84, 3 (2010), 394–421.
[42] Amit Lazarov, Ziv Ben-Zion, Dana Shamai, Daniel S Pine, and Yair Bar-Haim. 2018. Free viewing of sad and happy faces in depression: A potential target for attention bias modification. *Journal of affective disorders* 238 (2018), 94–100.
[43] Stephan Lewandowsky, Gilles E Gignac, and Klaus Oberauer. 2013. The role of conspiracist ideation and worldviews in predicting rejection of science. *PloS one* 8, 10 (2013), e75637.
[44] Geoffrey R Loftus. 1972. Eye fixations and recognition memory for pictures. *Cognitive psychology* 3, 4 (1972), 525–551.
[45] Shengfu Lu, Jiying Xu, Mi Li, Jia Xue, Xiaofeng Lu, Lei Feng, Bingbing Fu, Gang Wang, Ning Zhong, and Bin Hu. 2017. Attentional bias scores in patients with depression and effects of age: a controlled, eye-tracking study. *Journal of International Medical Research* 45, 5 (2017), 1518–1527.
[46] Franziska Marquart, Jörg Matthes, and Elisabeth Rapp. 2016. Selective exposure in the context of political advertising: A behavioral approach using eye-tracking methodology. *International Journal of Communication* 10 (2016), 20.
[47] Jennifer McCoy, Tahmina Rahman, and Murat Somer. 2018. Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62, 1 (2018), 16–42.
[48] Randall K Minas, Robert F Potter, Alan R Dennis, Valerie Bartelt, and Soyoung Bae. 2014. Putting on the thinking cap: using NeuroIS to understand information processing biases in virtual teams. *Journal of Management Information Systems* 30, 4 (2014), 49–82.
[49] Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper* 18-87 (2018).
[50] Atilla Alpay Nalcaci, Dilara Girgin, Semih Balki, Fatih Talay, Hasan Alp Boz, and Selim Balcisoy. 2019. Detection of Confirmation and Distinction Biases in Visual Analytics Systems.. In *TrustVis@ EuroVis*. 13–17.
[51] Efrat Nechushtai and Seth C Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (2019), 298–307.
[52] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.
[53] Alexander Nussbaumer, Katrien Verbert, Eva-Catherine Hillemann, Michael A Bedek, and Dietrich Albert. 2016. A framework for cognitive bias detection and feedback in a visual analytics environment. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. IEEE, 148–151.
[54] Łukasz Okruszek, Kirsty Dolan, Megan Lawrence, and Matteo Cella. 2017. The beat of social cognition: Exploring the role of heart rate variability as marker of mentalizing abilities. *Social neuroscience* 12, 5 (2017), 489–493.
[55] Margit E Oswald and Stefan Grosjean. 2004. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory* 79 (2004).
[56] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[57] Eric D Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. 2015. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 31–40.
[58] Michael Salter. 2016. Men's Rights or Men's Needs? Anti-Feminism in Australian Men's Health Promotion. *Canadian Journal of Women and the Law* 28, 1 (2016), 69–90.
[59] Desirée Schmuck, Miriam Tribastone, Jörg Matthes, Franziska Marquart, and Eva Maria Bergel. 2020. Avoiding the Other Side? An eye-tracking study of selective exposure and selective avoidance effects in response to political advertising. *Journal of Media Psychology* 32, 3 (2020), 158–164.
[60] Christoph Schneider, Markus Weinmann, and Jan vom Brocke. 2015. Choice architecture: Using fixation patterns to analyze the effects of form design on cognitive biases. In *Information systems and neuroscience*. Springer, 91–97.
[61] Milad Shokouhi, Ryen White, and Emine Yilmaz. 2015. Anchoring and adjustment in relevance estimation. In *Proceedings of the 38th International ACM SIGIR Conference on research and development in information retrieval*. 963–966.
[62] Namrata Srivastava, Joshua Newn, and Eduardo Velloso. 2018. Combining Low and Mid-level Gaze Features for Desktop Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–27.
[63] Michael Sülflow, Svenja Schäfer, and Stephan Winter. 2019. Selective attention in the news feed: An eye-tracking study on the perception and selection of political news posts on Facebook. *new media & society* 21, 1 (2019), 168–190.
[64] Jill M Swirsky and David Jason Angelone. 2016. Equality, empowerment, and choice: what does feminism mean to contemporary women? *Journal of Gender Studies* 25, 4 (2016), 445–460.
[65] Julian F Thayer, Fredrik Åhs, Mats Fredrikson, John J Sollers III, and Tor D Wager. 2012. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews* 36, 2 (2012), 747–756.
[66] Martin A Tolcott, F Freeman Marvin, and Paul E Lehner. 1989. Expert decision-making in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics* 19, 3 (1989), 606–615.
[67] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.
[68] Micah N Villarreal, Alexander J Kamrud, and Brett J Borghetti. 2019. Confirmation Bias Estimation from Electroencephalography with Machine Learning. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 63. SAGE Publications Sage CA: Los Angeles, CA, 73–77.
[69] Emily Wall, Arup Arcalgud, Kuhu Gupta, and Andrew Jo. 2019. A markov model of users' interactive behavior in scatterplots. In *2019 IEEE Visualization Conference (VIS)*. IEEE, 81–85.
[70] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. 2019. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *IFIP Conference on Human-Computer Interaction*. Springer, 555–575.
[71] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.
[72] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe* 27 (2017).
[73] Chris Wells, Katherine J Cramer, Michael W Wagner, German Alvarez, Lewis A Friedland, Dhavan V Shah, Leticia Bode, Stephanie Edgerly, Itay Gabay, and Charles Franklin. 2017. When we stop talking politics: The maintenance and closing of conversation in contentious times. *Journal of Communication* 67, 1 (2017), 131–157.
[74] Axel Westerwick, Benjamin K Johnson, and Silvia Knobloch-Westerwick. 2017. Confirmation biases in selective exposure to political online information: Source bias vs. content bias. *Communication Monographs* 84, 3 (2017), 343–364.
[75] Axel Westerwick, Daniel Sude, Melissa Robinson, and Silvia Knobloch-Westerwick. 2020. Peers versus pros: Confirmation bias in selective exposure to user-generated versus professional media messages and its consequences. *Mass Communication and Society* 23, 4 (2020), 510–536.
[76] Fabiana Zollo. 2019. Dealing with digital misinformation: a polarised context of narratives and tribes. *EFSA Journal* 17 (2019), e170720.
[77] Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. 2017. Debunking in a world of tribes. *PloS one* 12, 7 (2017), e0181821.
[78] Fabiana Zollo and Walter Quattrociocchi. 2018. Misinformation spreading on Facebook. In *Complex spreading phenomena in social systems*. Springer, 177–196.