Article

# Algorithmic or Human Source? Examining Relative Hostile Media Effect With a Transformer-Based Framework

Chenyan Jia [1],* and Ruibo Liu [2]

[1] Moody College of Communication, The University of Texas at Austin, USA; E-Mail: chenyanjia@utexas.edu
[2] Department of Computer Science, Dartmouth College, USA; E-Mail: ruibo.liu.gr@dartmouth.edu

* Corresponding author

## Abstract

The relative hostile media effect suggests that partisans tend to perceive the bias of slanted news differently depending on whether the news is slanted in favor of or against their sides. To explore the effect of an algorithmic vs. human source on hostile media perceptions, this study conducts a 3 (author attribution: human, algorithm, or human-assisted algorithm) × 3 (news attitude: pro-issue, neutral, or anti-issue) mixed factorial design online experiment (N = 511). This study uses a transformer-based adversarial network to auto-generate comparable news headlines. The framework was trained with a dataset of 364,986 news stories from 22 mainstream media outlets. The results show that the relative hostile media effect occurs when people read news headlines attributed to all types of authors. News attributed to a sole human source is perceived as more credible than news attributed to two algorithm-related sources. For anti-Trump news headlines, there exists an interaction effect between author attribution and issue partisanship while controlling for people's prior belief in machine heuristics. The difference of hostile media perceptions between the two partisan groups was relatively larger in anti-Trump news headlines compared with pro-Trump news headlines.

## Issue

This article is part of the issue "Algorithmic Systems in the Digital Society" edited by Sanne Kruikemeier (University of Amsterdam, The Netherlands), Sophie Boerman (University of Amsterdam, The Netherlands), and Nadine Bol (Tilburg University, The Netherlands).

## 1. Introduction

With advances in machine learning techniques and the growing availability of big data, algorithms have become widely adopted in news agencies around the world (Jia & Johnson, 2021). Automated journalism is defined as a form of news production that can automatically produce news stories with little human intervention beyond the initial programming phase (Carlson, 2015; Graefe, 2016; Tandoc et al., 2020). Although automated journalistic writing is mostly restricted to factual and data-driven topics such as sports, finance, crime, weather, and disaster reporting, it has also been applied to other domains such

as political news (Jia & Johnson, 2021; Wu, 2020). With the growing presence of automated journalism, this new technological affordance has altered how audiences consume and engage with news (Liu & Wei, 2019).

Increasing scholarly attention has been given to the perceptions of automated news (e.g., Graefe et al., 2018; Jia & Gwizdka, 2020; Wu, 2020). One recent meta-analysis shows that when reading the actual content written by humans and algorithms, people perceive no difference in terms of news credibility; however, people perceive news purportedly attributed to algorithms as slightly less credible than news attributed to humans (Graefe & Bohlken, 2020). Algorithmic author

attribution may reduce message credibility through an indirect pathway of source anthropomorphism because people prefer human rather than machine sources possibly due to the principle of similarity attraction (e.g., Byrne, 1997; Simons et al., 1970). Anthropomorphism is defined as the attribution of human traits, motivations, emotions, or behaviors to non-human and non-living entities (Airenti, 2015). Another explanation for why algorithmic sources are perceived as less credible than human sources is that people tend to be less familiar and knowledgeable with automation technologies (e.g., Haim & Graefe, 2017).

Despite the initial evidence of differences in source credibility between human and algorithmic sources, few studies have further examined whether machine source attribution can affect partisans' perceptions of news with an ideological slant. It remains unclear whether the manipulation of the source can affect people's perceptions of media bias, especially for partisans who are likely to fall prey to hostile media phenomena. Hostile media effect (HME) refers to the tendency for people who are highly involved in an issue to rate ostensibly neutral and balanced stories as biased due to their own biases (Arpan & Raney, 2003; Feldman, 2011; Giner-Sorolla & Chaiken, 1994; Vallone et al., 1985). The relative HME theory further argues that partisans tend to perceive the extent of bias of slanted news coverage differently depending on whether the news is slanted in favor of or against their points of view (Goldman & Mutz, 2011; Gunther et al., 2001). Previous studies have examined the relative HME by studying how partisans perceive news from sources with different levels of credibility (e.g., Arpan & Raney, 2003; Coe et al., 2008; Gunther & Liebhart, 2006). Gunther and Liebhart (2006), for instance, examine the influence of sources with different credibility (journalist vs. college student) on people's perceptions of bias. Their results suggest that more credible sources (journalist, large reach) yield more HME than lower credible sources (college student, small reach; Gunther & Liebhart, 2006). Few studies, however, have investigated how algorithmic sources will affect the relative HME compared with human sources. Will the attribution of an algorithmic source produce less HME than a human source because of its relatively lower credibility?

Given the increasing usage of AI in the online political context, it is especially important to examine whether algorithmic sources (as opposed to human sources) will increase or reduce relative HME. If an algorithmic cue can reduce the relative HME or perceived bias, it might increase partisans' exposure to cross-cutting information and help combat the extreme polarization (Jia & Johnson, 2021). In order to examine the effect of algorithmic sources on HME, this study conducts a 3 (author attribution: human, algorithm, or human-assisted algorithm) × 3 (news attitude: pro-issue, neutral, or anti-issue) mixed factorial design online experiment (N = 511). Using the computational method, this study adopts a transformer-based adversarial network to generate

comparable stimuli. The framework was trained with a dataset of 364,986 news stories from 22 mainstream media outlets (Liu, Jia, & Vosoughi, 2021). The present research answers the following overarching questions: (a) Will stories purportedly written by humans produce different relative HME compared with those by algorithms? (b) How will source credibility (human vs. algorithmic sources) affect relative HME?

## 2. Literature Review

### 2.1. Relative Hostile Media Effect

HME refers to the tendency for partisans (i.e., individuals with a strong preexisting political stance) to perceive neutral media coverage as biased against their sides (e.g., Giner-Sorolla & Chaiken, 1994). The original HME theory assumes that news stories are balanced (Feldman, 2011). The relative HME theory expands the assumption of the original HME by making it applicable to news that is slanted rather than balanced (Gunther et al., 2001). More specifically, the relative HMEs theory suggests that partisans tend to perceive the same media content differently and perceive less bias in the news coverage leaning toward their views than their opponents (Feldman, 2011; Gunther & Chia, 2001). The HME theory has been tested in numerous contexts through both experimental and survey methods (Feldman, 2017; Perloff, 2015). A recent meta-analysis of 34 HME studies has shown that a considerable number of empirical studies have provided widespread evidence of HME (Hansen & Kim, 2011).

Researchers have attempted to provide multiple explanations for why HME manifests itself. One explanation is the idea of the selective process (or message-processing mechanisms; Feldman, 2011, 2017; Giner-Sorolla & Chaiken, 1994; Gunther & Liebhart, 2006). Partisans often selectively recall, categorize, or use different standard mechanisms to process unfavorable or attitude-challenging content (Hansen & Kim, 2011). The second core factor that may lead to HME is the level of involvement. Early HME studies often suggest that hostile media phenomenon is limited to partisans who have strong issue involvement (e.g., Giner-Sorolla & Chaiken, 1994; Vallone et al., 1985). Recent studies, however, view issue involvement as a moderator of HME (Perloff, 2015).

Another key factor that might explain the HME is source credibility (Hansen & Kim, 2011). Arpan and Raney (2003) suggest that the credibility ratings of news sources may affect hostile media perceptions (HMP). People's expectations about the media outlet affect the perceived hostility of the media (Giner-Sorolla & Chaiken, 1994). Past work indicates that people's prior beliefs about the source credibility (or a related concept "trust") give rise to biased processing of the content (Baum & Gussin, 2007). Partisans often perceive news sources providing confirmatory information as more credible than those that do not (Baum & Gussin, 2007). Some studies

have examined how source credibility affects HME by examining home-town newspapers vs. rival-town newspapers (Arpan & Raney, 2003), college students vs. journalists (Gunther & Liebhart, 2006), different cable television news programs (CNN, FOX, The Daily Show; Coe et al., 2008). Other studies have also examined relative HME by varying general characteristics of news coverage (circulation size, type of ownership; Gunther, 1992).

## 2.2. Automated Journalism

Very few studies have examined whether human sources and algorithmic sources will yield different relative HME. Many scholars have conducted empirical research on perceptions of automated news by examining both the actual content of automated news (e.g., Clerwall, 2014; Graefe et al., 2018; Haim & Graefe, 2017; Jia & Gwizdka, 2020; Wu, 2020) and the effects of machine vs. human source attribution (e.g., Jung et al., 2017; Tandoc et al., 2020; Waddell, 2019). In terms of perceptions of credibility, studies examining actual content produced by algorithms as opposed to humans yield different results from studies focusing exclusively on the effect of source attribution (e.g., Graefe & Bohlken, 2020; Jia, 2020; Waddell, 2018). One recent meta-analysis including 12 studies on automated news shows no difference in readers' perceptions of credibility when reading the actual content written by humans and algorithms (Graefe & Bohlken, 2020). In terms of the effect of source, however, findings of the meta-analysis revealed that people perceive news purportedly attributed to algorithms as slightly less credible than news attributed to humans (Graefe & Bohlken, 2020).

To understand the effect of machine sources as opposed to human sources, many recent studies examine the source attribution while controlling for the content (e.g., Jung et al., 2017; Tandoc et al., 2020; Waddell, 2019). Results are mixed on whether news purportedly written by a machine source is more or less credible. Some studies found that news attributed to a machine author is more credible than news attributed to a human author, especially for news that requires more information processing (Liu & Wei, 2019; Waddell, 2019). Others found no main difference in the perceived source credibility between news attributed to algorithmic and human authors (e.g., Tandoc et al., 2020). Overall, most previous studies suggest that news attributed to a human author is perceived as more credible than news attributed to an algorithmic author (Graefe & Bohlken, 2020; Jia & Johnson, 2021).

Very few studies, however, have further investigated whether the difference in source credibility (human vs. algorithm) will have an impact on relative HME. Therefore, this study aims to fill in the gap by examining whether HME also occurs in news stories attributed to algorithms. This work proposes the authorship (humans, algorithms, or human-assisted algorithms) of news stories as a novel source cue and examines how differ-

ent sources affect issue partisans' HMP. Issue partisans are people who hold strong and even extreme attitudes toward an issue, especially a political issue (Feldman, 2011). Previous work suggests that partisans in favor of one issue often perceive anti-issue news as relatively more biased than partisans on the opposing side, regardless of the news source (e.g., Arpan & Raney, 2003). Pro-issue news refers to the news in favor of one issue whereas anti-issue news refers to the news standing against one issue. Adding to previous literature, this study predicts the following hypotheses:

H1: For news headlines purportedly written by (a) humans, (b) algorithms, and (c) human-assisted algorithms, partisans on the supporting side of an issue will perceive the anti-issue news as relatively more biased than partisans on the opposing side.

H2: For news headlines purportedly written by (a) humans, (b) algorithms, and (c) human-assisted algorithms, partisans on the opposing side of an issue will perceive the pro-issue news as relatively more biased than partisans on the supporting side.

## 2.3. Source Credibility

Source credibility was initially used to measure how the characteristics of speakers influence the receiver's acceptance of a message (Hovland et al., 1953). Factors such as the speaker's expertise, truthfulness, and motivation to tell the truth are major characteristics to determine source credibility (e.g., Hovland et al., 1953; Perloff, 2015). The concept of credibility is related to different theoretical concepts including trust and fairness (Engelke et al., 2019). People are inclined to judge news stories from traditional mainstream media as more credible than those from social media because they often employ journalistic values such as trustworthiness, fairness, professionalism, and balance in assessing credibility (Johnson & Kaye, 2013; Yamamoto et al., 2016). People also tend to perceive attitude-consistent information as more credible and trustworthy than that challenges their beliefs (Metzger & Flanagin, 2015).

A large body of literature suggests the influence of source credibility on perceptions of media bias (e.g., Arpan & Raney, 2003; Gunther & Liebhart, 2006). For instance, Gunther and Liebhart (2006) found that more traditionally credible sources (i.e., journalists over students, large over small reach) facilitate more relative HME. The majority of past studies on automated journalism suggest that news purportedly written by a human author is perceived as more credible than that by an algorithmic author (e.g., Graefe & Bohlken, 2020; Waddell, 2019). Adding to previous literature, this study also predicts that people will perceive human sources as more credible, and thus yield greater HME in news attributed to human sources rather than news attributed to algorithmic sources:

H3: People will perceive human sources as more credible than algorithms sources.

H4: Relative HME will be greater for news headlines purportedly written by humans than news headlines purportedly written by algorithms or human-assisted algorithms.

Readers who process information heuristically often rely on the reputation of heuristic cues and consider familiar sources more credible (Metzger & Flanagin, 2015; Metzger et al., 2010). Several automated journalism studies have shown that audiences have relatively low familiarity and lack of knowledge with automation technologies (e.g., Haim & Graefe, 2017) and thus may perceive algorithm attributed news differently from human attributed news in terms of credibility (e.g., Clerwall, 2014; Jung et al., 2017; Waddell, 2019). Readers may rely more on heuristic cues such as source credibility to make judgments rather than their own issue attitudes because of their low familiarity with the underlying mechanism of algorithms (Haim & Graefe, 2017). One previous study found that source credibility partially mediates the influence of issue partisanship on peoples' selective exposure to gun stories (Jia & Johnson, 2021). Another study also found that trust in source mediates people's perceptions of algorithmic products (Shin, 2020). Adding to past work, this present study predicts a mediating effect of source credibility on people's HMP:

H5: Source credibility will mediate the influence of issue partisanship on people's HMP.

## 3. Method

### 3.1. Experimental Design

The present study adopted a 3 (author attribution: human, algorithm, or human-assisted algorithm) × 3 (news attitude: pro-issue, neutral, or anti-issue) mixed factorial design online experiment. An online experiment ($N$ = 511) embedded in Qualtrics was conducted in January 2021. Author attribution was a between-subjects variable whereas news attitude was a within-subjects variable. Participants were randomly assigned to read news headlines purportedly written by a human author ($n$ = 168), algorithm ($n$ = 168), or human-assisted algorithm ($n$ = 175). Each participant was asked to read 15 news headlines about Donald Trump. The order of headlines was randomized.

### 3.2. Procedures

Before the experiment, participants were asked to report their political attitude, party affiliation, and attitude towards Donald Trump. Participants also need to answer several questions about source familiarity and source credibility for both human authors and algorithmic authors. Then, participants were randomly assigned to read news headlines purportedly written by human, algorithm, or human-assisted algorithm. After participants read each headline, they were asked to rate their perceived bias and credibility of news stories. Given the important role of author attribution in this experimental design, participants were asked if they could recall the author listed on the byline (adapted from Jia & Johnson, 2021; Waddell, 2019). Two attention checks were embedded in the experiment to exclude careless responses. Participants were asked to select point three on the first attention check question. In the second attention check question, participants were asked to add three to the first number they selected and to use the result as the answer to the second question.

### 3.3. Participants

For both the pre-test and the main experiment, participants were recruited from CloudResearch (formerly known as TurkPrime) which is an advanced online crowd-sourcing platform for behavioral science data collection (Litman et al., 2017). Participants were all from the United States and above 18 years old. Each participant was required to have a HIT approval rate greater than 95%. Participants in both pre-test and main experiment were paid 75 cents for their participation. After ruling out repeated IP addresses, incomplete answers, and subjects who failed both attention checks, 511 participants remained in the main experiment. The average age of participants was 41.10 years old ($SD$ = 12.77, $Median$ = 39). More than half of the participants (54%) were male, 45.8% were female, and 0.2% of participants chose other. Participants have received 16.27 years of education on average ($SD$ = 2.30, $Median$ = 16). The majority of participants are White (73.6%), followed by 11.7% Asian/Pacific Islander, 7.8% Black/African American, 3.3% Hispanic/Latino/Latina, 2.7% Other/Multi-Racial, and 0.8% participants preferred not to respond to the race question. About a half of the participants (48.1%) self-identified as Democrats, 24.1% were Independent, and 27.8% self-identified as Republicans.

### 3.4. Stimuli and Computational Method

In order to guarantee the stimuli were comparable, this study used a computer-assisted method, the current state-of-the-art controllable headline generation model (Liu, Jia, & Vosoughi, 2021), to generate news headlines with different political ideologies (i.e., liberal, neutral, and conservative). Specifically, the model consists of two main modules: the polarity detection module and the polarity flipper module. The polarity detector leverages the self-attention mechanism of the Transformer framework to score the polarity of different spans of the text and outputs the biased part of the text. The polarity flipper only flips the detected biased content through

an adversarial network, with the preservation of semantic consistency (Liu, Jia, & Vosoughi, 2021). The model was trained on a dataset of 364,986 news stories from 22 mainstream media outlets (CNN, NYTimes, Fox, WSJ, etc.). News articles were collected from Allsides and Media Cloud. Each news story from Allsides was labeled with a political polarity label by an editing expert. News articles from Media Cloud were assigned with ideological polarity labels according to the polarity of media outlets using the rationale developed by Pew Research Center (Pew Research, 2014). All news articles were collected from June 2012 to May 2019.

We first selected eight news headlines about the 45th President of the United States, Donald Trump, from our dataset (four from liberal media outlets and four from conservative media outlets). News headlines were selected as stimuli because our model performed better in flipping and neutralizing headlines rather than body text. This can be explained by the fact that issue attitudes are oftentimes more obvious in the headlines of strategic news stories compared with the body text (Liu, Wang, et al., 2021). Using the transformer-based framework, this study automatically generates corresponding headlines either with the opposite polarity or being neutral. Some of the machine-generated headlines may be slightly ungrammatical due to the mechanism of machine learning algorithms (e.g., *Trump Said That Shouldn't Matter Those State of the Union Ratings*). However, we intentionally did not include an additional human editing process because the post-editing procedure of machine-generated texts may introduce more potential bias (Biswas & Rajan, 2020). All stimuli were generated by one unified transformer-based framework and thus have an overall consistent performance on the grammatical level. In order to avoid the potential influence of ungrammatical languages, this study used multiple sets of headlines to test the robustness of the experiment design.

A pre-test (*N* = 90) was conducted to examine the issue attitudes of these 24 headlines. Participants in the pre-test were asked to read 24 headlines and answer whether the news headline was strictly neutral or biased in favor of one side or the other followed by the 11-point scale with −5 indicating strongly biased against Trump, and +5 indicating strongly biased in favor of Trump, 0 indicating strictly neutral. Several one-sample T-tests (test value = 0) were conducted to test whether the stimuli were biased in the direction they were designed to be. Based on the results of the pre-test, 3 sets of headlines were excluded in the main experiment because they were not statistically significant in the direction they were designed to be. In total, 15 headlines were chosen as final stimuli of the main experiment (5 anti-Trump, 5 neutral, 5 pro-Trump), as shown in Table 1.

This study used Photoshop to make every stimulus looks like a screenshot from the same fictional news site. The bylines of articles were under each headline, following with the published time and three social media sharing buttons. In the byline, it either shows "by an automated journalism algorithm," "by staff reporter Jim Richard," or "by an automated journalism algorithm and staff reporter Jim Richard."

### 3.5. Measures

#### 3.5.1. Issue Partisanship

Issue partisanship (*M* = 1.60, *SD* = 0.59) was measured by asking to what extent they support or oppose Trump. Responses were recorded on the 11-point scale ranging from −5 (*strongly support Trump*), 0 (*strictly neutral*), and +5 (*strongly oppose Trump*; adapted from

**Table 1.** News headlines stimuli.

| Anti-Trump | Neutral | Pro-Trump |
|---|---|---|
| Donald Trump Lied About His State of The Union Ratings | Trump Said That Shouldn't Matter Those State of The Union Ratings | Trump Claims Highest His State of The Union Ratings |
| Trump Denies Asking Ex-FBI Director Comey to Drop Flynn Investigation | Trump Says: "I Never Asked Comey to Stop Investigating Flynn'' | Trump Never Asked Ex-FBI Director Comey to Stop Investigating Flynn |
| Trump Administration Considers Tearing Families Apart in New Immigration Crackdown | Trump Immigration Plan Provides Path to Citizenship for Millions of Immigrants Illegally | Trump Offers Dreamers A Path to Citizenship, Tough on Other Immigrants |
| Trump Threatens to Abandon Puerto Rico Recovery Effort | Senate Narrowly Approves Budget, Paving Way for Tax Reform | Senate Approves Budget in Crucial Step for Trump's Tax Overhaul |
| Trump Administration Approves Plan to Separate Families at Border | Trump's New HHS Office Will Protect Health Care Workers Who Violate Abortion | Trump Administration Creates New Religious Protections for Health Care Workers |

Notes: The bias ratings of anti-Trump stimuli are significantly lower than value 0 whereas ratings of pro-Trump stimuli are all significantly higher than value 0; the ratings of neutral headlines are not significantly different from value 0.

Feldman, 2011). Participants were then categorized into three groups using the cutoff value 0. Participants who selected 0 (strictly neutral) were classified as non-partisans and were excluded in further analysis. Participants scoring below 0 were classified as pro-Trump partisans ($n$ = 148) whereas those scoring above 0 were classified as anti-trump partisans ($n$ = 336).

### 3.5.2. Hostile Media Perception

The measure of perceived bias or slant in the news headlines was adapted from previous research on HMP (Giner-Sorolla & Chaiken, 1994; Gunther & Schmitt, 2004; Feldman, 2011). Participants were asked "Would you say that the above news headline was strictly neutral, or was it biased in favor of one side or the other?" followed by the 11-point scale with –5, *strongly biased against Trump*, and +5, *strongly biased in favor of Trump*. Two additional items asked participants to rate what percentage of the news headline was unfavorable and favorable, respectively, toward the focal news issue on 11-point scales ranging from 0 to 100%. Both items were converted to a –5 to +5 response scale. All three items were then averaged to form a scale, where positive scores represent the headline is perceived as biased favorable toward Trump and negative scores a bias unfavorable toward Trump. Three items were highly correlated and can be averaged to form a reliable index ($M$ = 1.16, $SD$ = 2.31, Cronbach's $\alpha$ = 0.78). This study also measured the HMP of authorship because the manipulation of the authorship is a key variable. Participants need to answer, "Would you say that the author of the above news headline was strictly neutral, or was it biased in favor of one side or the other?" on a –5 to +5 response scale (adapted from Gunther & Schmitt, 2004).

### 3.5.3. Source Credibility and Message Credibility

Credibility was measured as a multidimensional construct consisting of believability, fairness, accuracy, depth of information, and authenticity (adapted from Gaziano & McGrath, 1986; Metzger et al., 2003; Newhagen & Nass, 1989) on 7-point scales. Participants were asked to rate source credibility based on their expectations or previous experiences with human, human-assisted algorithm, and algorithmic authors. After reading each news headline, participants were asked to rate message credibility. This study reverse coded the authenticity item ("the story written by humans or algorithms is not authentic"). Five items were highly correlated and can be averaged into one measure ($M$ = 3.99, $SD$ = 1.32, Cronbach's $\alpha$ = 0.92).

### 3.5.4. Prior Belief in Machine Heuristics

Participants' prior beliefs in the machine heuristic were measured using four bipolar items "harmful/beneficial," "unethical/ethical," "unfavorable/favorable,'' "unneces-

sary/necessary." Participants were asked to rate the questions "What is your view on using machine learning software to replace or augment human journalists?" on 7-point scales. One item "unfavorable/favorable" was reverse coded. Four items were highly correlated and can be averaged into one measure ($M$ = 3.15, $SD$ = 1.42, Cronbach's $\alpha$ = 0.90).

### 3.6. Manipulation Check

To ensure the experimental manipulation was effective, participants answered a manipulation check to rate their perceived source anthropomorphism of the listed author(s) on four 7-point scales. Four semantic differential items "fake/natural," "unconscious/conscious," "artificial/life-like," and "mechanical/organic" were adapted from prior research to measure the perceived source anthropomorphism (Bartneck et al., 2007; Jia & Johnson, 2021). Four items were highly correlated and can be averaged to form a reliable index (Cronbach's $\alpha$ = 0.94). One-way Analysis of Variance (ANOVA) was conducted to compare the difference of the perceived source anthropomorphism among the three groups. The source anthropomorphism ($M$ = 4.44, $SD$ = 1.47) rated by participants ($n$ = 168) who were assigned to the human group was significantly higher than the source anthropomorphism ($M$ = 4.13, $SD$ = 1.34) rated by participants ($n$ = 175) who were assigned to the human-assisted algorithm group and that of ($M$ = 3.02, $SD$ = 1.44) the algorithmic author group ($n$ = 168), $F(2, 508)$ = 46.79, $p$ < 0.001, which showed the manipulation was successful.

## 4. Results

A two-way repeated measures ANOVA was conducted to test H1 and H2. H1 predicted that for anti-Trump news headlines purportedly written by all types of authors, the pro-Trump group would perceive more bias than the anti-Trump group. Analysis showed that there was a significant difference between the anti-Trump and the pro-Trump group, $F(1,478)$ = 59.45, $p$ < 0.001, $\eta^2$ = 0.11. For anti-Trump news attributed to each author type, the pro-Trump group perceived significantly more bias than the anti-Trump group, $p$ < 0.001. H1 was supported.

H2 predicted that for the pro-Trump news headlines written by all types of authors, anti-Trump partisans would perceive more bias than the pro-Trump group. Analysis showed that there was a significant difference between two groups, $F(1,478)$ = 17.95, $p$ < 0.001, $\eta^2$ = .04. The direction was as expected in each condition, as shown in Table 2. Therefore, H1 and H2 were supported.

H3 predicted that people would perceive human sources as more credible than algorithms sources. A one-way ANOVA was conducted to examine whether there existed significant differences in source credibility between algorithm-related authors and human authors. Results showed that human sources ($M$ = 4.19, $SD$ = 1.24) were perceived as significantly more credible

**Table 2.** Partisan group HMP means in three authorship conditions.

| | Algorithm Author | | Combined Author | | Human Author | |
|---|---|---|---|---|---|---|
| | Anti-Trump group | Pro-Trump group | Anti-Trump group | Pro-Trump group | Anti-Trump group | Pro-Trump group |
| **Anti-Trump News** | | | | | | |
| *M* | −1.95(0.13) | −3.02(0.19) | −2.13(0.18) | −3.25(0.21) | −1.98(0.13) | −2.91(0.18) |
| *n* | 106 | 53 | 128 | 40 | 102 | 55 |
| | | | | | | |
| **Neutral News** | | | | | | |
| *M* | 0.51(0.11) | −0.37(0.15) | 0.30(0.10) | −0.55(0.17) | 0.38(0.11) | −0.10(0.15) |
| *n* | 106 | 53 | 128 | 40 | 102 | 55 |
| | | | | | | |
| **Pro-Trump News** | | | | | | |
| *M* | 1.51(0.12) | 0.73(0.17) | 1.38(0.11) | 0.76(0.20) | 1.40(0.12) | 1.22(0.17) |
| *n* | 106 | 53 | 128 | 40 | 102 | 55 |

Note: Means for each group were presented as marginal means (with standard errors in parentheses).

than both the algorithm author ($M = 3.75$, $SD = 1.35$) and the combined author ($M = 4.07$, $SD = 1.28$), $F_{(2,508)} = 5.16$, $p < 0.01$. Specifically, pure algorithmic sources received the lowest credibility score. Therefore, H3 was supported.

H4 predicted that the relative HME would be greater for news headlines purportedly written by humans compared with news headlines purportedly written by algorithms or human-assisted algorithms. Two-way repeated measures ANOVA showed that there was no significant main effect observed for source attribution in both anti-Trump, $F_{(2,478)} = 1.18$, $p = 0.31$, $\eta^2 = 0.01$, and pro-Trump headlines, $F_{(2,478)} = 1.44$, $p = 0.24$, $\eta^2 = 0.01$. The interaction between source attribution and issue attitudes was not significant for both anti-Trump, $p = 0.85$, and pro-Trump news headlines, $p = 0.12$. Therefore, H4 was not supported. In fact, in both pro- and anti-Trump news headlines, human authors produced smaller relative HME than algorithm-related authors, but the difference is not statistically different, as shown in Figure 1.

Several repeated measures ANCOVAs were conducted as additional analyses. Results show that while controlling for people's prior belief in machine heuristics, a marginally significant interaction effect between

source attribution and issue attitudes was detected for anti-Trump news headlines, $p = 0.052$. There existed no significant interaction effect between source attribution and issue partisanship for pro-Trump news headlines, $p = 0.11$ and neutral headlines, $p = 0.41$ after controlling for people's prior belief in machine heuristics.

H5 predicted that source credibility would mediate the influence of issue partisanship on the perceived bias. To test this hypothesis, a mediation model was run by using the PROCESS macro (Hayes, 2013) based on nonparametric bootstrapping with 1,000 simulations and 95% confidence intervals (CIs). For anti-Trump news headlines, the effect of issue partisanship on HMP was partially mediated via source credibility. Ninety-five percent CIs for indirect, direct, and total effects did not include zero, which means all these effects were significant. As Figure 2 illustrates, for anti-Trump news, the indirect effect was $a_1 * b_1 = 0.016$, CI = [0.001, 0.03]. The direct effect was $c_1 = 0.13$, CI = [0.095, 0.17]. The total effect was 0.15, CI = [0.11, 0.19]. Thus, H5 was supported for anti-Trump news headlines. For pro-Trump news headlines, a mediation effect of source credibility did not occur. The indirect effect was not significant, CI = [−0.03, 0.00]. Therefore, H5 was partially supported.
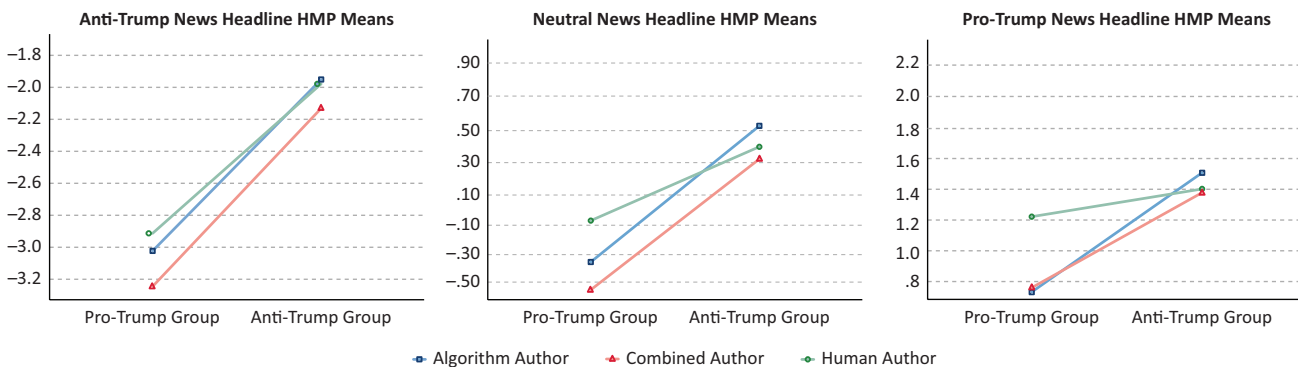


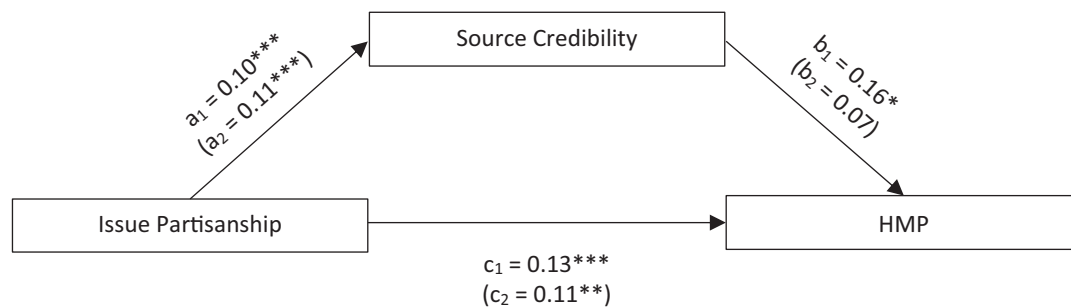**Figure 1.** Perceived headline HMP means.

**Figure 2.** Mediation analysis of anti-Trump (pro-Trump) news headlines. Notes: Two mediation models were presented in the figure; source credibility partially mediated the influence of issue attitudes on HMP for anti-Trump news headlines; *$p < 0.05$, **$p < 0.01$ *** $p < 0.001$.

## 5. Discussion

### 5.1. Theoretical Contribution

In recent years, algorithms become ubiquitous in the contemporary media environment (Thurman et al., 2017). Our work adds to the previous literature in both fields of political communication and human-computer interaction. Previous works have studied how different types of traditional news sources such as different cable television channels (Coe et al., 2008), journalists vs. college students (Gunther & Liebhart, 2006) affected the relative HME. Our study introduced a novel type of source (algorithmic sources) and shed light on new trends in the era of an algorithms-driven society.

This study contributed to both automated journalism and HME literature. First, the results of this study showed that relative HME occurred in all types of author attributions (human, algorithm, and combined authors). This work revealed that partisans tend to perceive the bias of slanted news coverage differently depending on their own political ideology even when the authors are algorithms. This result was consistent with previous research that relative HME existed regardless of the news source (e.g., Arpan & Raney, 2003). This work also found that news attributed to a sole human source was perceived as more credible than two algorithm-related sources. News attributed to a sole algorithmic author was rated as the least credible among three types of declared authors. This result was consistent with most previous studies (see Graefe & Bohlken, 2020; Jia & Johnson, 2021) but contradicted Waddell's (2019) findings. Waddell (2019) suggests that messages attributed to human authors are perceived as the least credible than those attributed to machine and combined sources. One possible explanation is that Waddell (2019) not only attributed authorship (human, algorithm, combined) but also attributed media outlets as another source to indicate the ideological slant of the stimuli (i.e., Fox News or MSNBC). Therefore, the credibility of media outlets may affect participants' perceived credibility of messages as well. Also, the difference of perceived message credibility among news attributed to three types of authors is marginal significant in Waddell's (2019) study.

This study predicted that news headlines attributed to human authors would evoke a larger relative HME than news attributed to algorithmic sources because one previous study suggests that more traditionally credible sources yield larger HME than lower credible sources (Gunther & Liebhart, 2006). This hypothesis was not supported. In fact, for both pro- and anti-Trump news headlines, human authors produced smaller relative HME than algorithm-related authors, but the difference was not statistically different. An interesting pattern was that when controlling for people's prior belief in machine heuristics, an interaction effect was detected between source attribution and issue partisanship. It is worth noting that such interaction effect was only significant for anti-Trump news headlines. With that said, for anti-Trump news headlines, news attributed to a sole algorithmic author (low credibility) yields larger relative HME than news attributed to a sole human author (high credibility) when controlling for people's prior belief in machine heuristics. As a matter of fact, even Gunther and Liebhart (2006) acknowledged that it is possible that partisans simply considered the student a more trustworthy source than the journalist. Future studies can further explore whether the pattern of low credible source yield larger relative HME stays true or not in other realms.

Another interesting finding of this study is that the difference of perceived bias between two partisan groups was relatively larger in anti-Trump news headlines compared with pro-Trump news headlines. The pro-Trump group perceived anti-Trump news headlines as much more biased than the anti-Trump group did. This is not surprising because Trump supporters are less likely to interact with outgroups (Pettigrew, 2017), and thus may perceive attitude-challenging information as very biased.

This study further investigated how source credibility affects people's perceived bias and found that source credibility partially mediates the influence of issue partisanship on people's perceived bias for anti-Trump news. This finding was interesting as it suggested that when partisans read news headlines, source credibility plays an important role in hostile media perceptions. This study posits a possible theoretical model to predict the perceived bias through source credibility.

Another contribution of this study is the implementation of a computer-assisted method to generate comparable news headlines. Earlier studies often use news content from different media outlets (e.g., *The New York Times* and *The Wall Street Journal*) to manually manipulate the select news articles or headlines with different political ideologies (e.g., Van Duyn & Collier, 2019). Although the political ideology of news content can be manipulated by pre-testing the stimuli, it is inevitable to include potential bias caused by different writing quality or source credibility (Liu, Jia, & Vosoughi, 2021). Our study, however, proposes a novel approach that is capable of flipping the ideology without shifting semantic meaning and readability of news articles.

### 5.2. Limitations

Despite these contributions, this study still has certain limitations. First, this study chose Trump-related news as a partisan issue because Trump was one of the most salient topics in our dataset. Our transformer-based framework was trained using the media coverage data from June 2012 to May 2019. Therefore, the stimuli were somewhat outdated by the time when the experiment was conducted. Future studies can choose more conventional political topics such as gun control, abortion, immigration, and gay marriage (Knobloch-Westerwick et al., 2017; Wojcieszak, 2019). Furthermore, some of the stimuli were not as readable as real news headlines due to the auto-generation process. The topic of the stimuli was not always controlled in each set because we prioritized the fact whether the ideological direction of stimuli was as we expected from the pre-test.

Second, the majority of our sample self-reported as Democrats, which also leads to the imbalanced number of two-issue partisan groups (pro-Trump and anti-Trump). Even though CloudResearch overcomes many limitations of Amazon Mechanical Turk, it still cannot represent the overall population as participants self-select studies to participate (Litman et al., 2017). Future research can recruit more representative samples especially in terms of pollical ideology.

Third, both issue partisanship and source credibility were not manipulated experimentally, which limits the plausibility of mediation models. If the mediator is measured rather than manipulated, one cannot exclude the possibility that a confounding variable may influence the relationship (Spencer et al., 2005). As one recent review of mediation analysis suggests that elaborate statistical techniques for testing mediation cannot overcome the flaws of inadequate research design (Chan et al., 2020).

### 5.3. Conclusion

As technology diffuses, the importance of examining how algorithmic source attribution will reduce or increase relative HME is of importance because such study bears implications to media effects studies as well as the media industry. AI research in the political communication area is still at a nascent stage. Some scholars contend that people's perceptions of news bias may be attenuated when news is attributed to a machine author (e.g., Waddell, 2019; Wang, 2021) because AI is often perceived as fair, objective, unbiased, and with less political agenda (Gillespie, 2014). This study, along with many others, found that such positive perceptions of machine neutrality may not always be true. Results of this study showed that the relative HME still occurs when people read news headlines attributed to algorithmic authors. In fact, news headlines attributed to algorithmic authors exhibited larger relative HME than those attributed to human authors in terms of anti-Trump news while controlling for people's prior belief in machine heuristics. The current study sheds light on a better understanding of the role of machine cues in the political context.

### Conflict of Interests

The authors declare no conflict of interests.

### References

Airenti, G. (2015). The cognitive bases of anthropomorphism: From relatedness to empathy. *International Journal of Social Robotics*, *7*(1), 117–127. https://doi.org/10.1007/s12369-014-0263-x

Arpan, L. M., & Raney, A. A. (2003). An experimental investigation of news source and the hostile media effect. *Journalism & Mass Communication Quarterly*, *80*(2), 265–281. https://doi.org/10.1177/107769900308000203

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The influence of people's culture and prior experiences with Aibo on their attitude towards robots. *AI and Society*, *21*, 217–230. https://doi.org/10.1007/s00146-006-0052-7

Baum, M. A., & Gussin, P. (2007). In the eye of the beholder: How information shortcuts shape individual perceptions of bias in the media. *Quarterly Journal of Political Science*, *3*, 1–31. http://doi.org/10.1561/100.00007010

Biswas, S., & Rajan, H. (2020). Do the machine learning models on a crowd sourced platform exhibit bias? An empirical study on model fairness. In P. Devanbu (Ed.), *Proceedings of the 28th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering* (pp. 642–653). ACM. https://doi.org/10.1145/3368089.3409704

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships*, *14*, 417–431. https://doi.org/10.1177/0265407597143008

Carlson, M. (2015). The robotic reporter. *Digital Jour-*

![](data:image/png;base64,...)
*nalism*, *3*(3), 416–431. https://doi.org/10.1080/2167 0811.2014.976412

Chan, M., Hu, P., & K. F. Mak, M. (2020). Mediation analysis and warranted inferences in media and communication research: Examining research design in communication journals from 1996 to 2017. *Journalism & Mass Communication Quarterly*. Advance online publication. https://doi.org/10.1177/10776 99020961519

Clerwall, C. (2014). Enter the robot journalist. *Journalism Practice*, *5*, 519–531. https://doi.org/10.1080/17512786.2014.883116

Coe, K., Tewksbury, D., Bond, B. J., Drogos, K. L., Porter, R. W., Yahn, A., & Zhang, Y. (2008). Hostile news: Partisan use and perceptions of cable news programming. *Journal of Communication*, *58*(2), 201–219. https://doi.org/10.1111/j.1460-2466.2008.00381.x

Engelke, K. M., Hase, V., & Wintterlin, F. (2019). On measuring trust and distrust in journalism: Reflection of the status quo and suggestions for the road ahead. *Journal of Trust Research*, *9*(1), 66–86. https://doi.org/10.1080/21515581.2019.1588741

Feldman, L. (2011). Partisan differences in opinionated news perceptions: A test of the hostile media effect. *Political Behavior*, *33*(3), 407–432.

Feldman, L. (2017). The hostile media effect. In K. Kenski & K. H. Jamieson (Eds.), *The Oxford handbook of political communication* (pp. 557–568). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199793471.013.011_update_001

Gaziano, C., & McGrath, K. (1986). Measuring the concept of credibility. *Journalism Quarterly*, *63*(3), 451–462. https://doi.org/10.1177/107769908606300301

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media technologies* (pp. 167–194). MIT Press. https://10.7551/mitpress/9780262525374.003.0009

Giner-Sorolla, R., & Chaiken, S. (1994). The causes of hostile media judgments. *Journal of Experimental Social Psychology*, *30*(2), 165–180. https://doi.org/10.1006/jesp.1994.1008

Goldman, S. K., & Mutz, D. C. (2011). The friendly media phenomenon: A cross-national analysis of crosscutting exposure. *Political Communication*, *28*(1), 42–66. https://doi.org/10.1080/10584609.2010.544 280

Graefe, A. (2016). *Guide to automated journalism*. Tow Center for Digital Journalism, Columbia University. https://doi.org/10.7916/D80G3XDJ

Graefe, A., & Bohlken, N. (2020). Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication*, *8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Graefe, A., Haim, M., Haarmann, B., & Brosius, H.-B. (2018). Readers' perception of computer-generated news: Credibility, expertise, and readability. *Journalism*, *19*(5), 595–610. https://doi.org/10.1177/1464884916641269

Gunther, A. (1992). Biased press or biased public? Attitudes toward media coverage of social groups. *The Public Opinion Quarterly*, *56*(2), 147–167. www.jstor.org/stable/2749167

Gunther, A. C., & Chia, S. C. Y. (2001). Predicting pluralistic ignorance: The hostile media perception and its consequences. *Journalism & Mass Communication Quarterly*, *78*(4), 688–701. https://doi.org/10.1177/107769900107800405

Gunther, A. C., Christen, C. T., Liebhart, J. L., & Chih-Yun Chia, S. (2001). Congenial public, contrary press, and biased estimates of the climate of opinion. *Public Opinion Quarterly*, *65*(3), 295–320. https://doi.org/10.1086/322846

Gunther, A. C., & Liebhart, J. L. (2006). Broad reach or biased source? Decomposing the hostile media effect. *Journal of Communication*, *56*(3), 449–466. https://doi.org/10.1111/j.1460-2466.2006.00295.x

Gunther, A. C., & Schmitt, K. (2004). Mapping boundaries of the hostile media effect. *Journal of Communication*, *54*(1), 55–70. https://doi.org/10.1111/j.1460-2466.2004.tb02613.x

Haim, M., & Graefe, A. (2017). Automated news: Better than expected? *Digital Journalism*, *5*, 1044–1059. https://doi.org/10.1080/21670811.2017.1345643

Hansen, G. J., & Kim, H. (2011). Is the media biased against me? A meta-analysis of the hostile media effect research. *Communication Research Reports*, *28*(2), 169–179. https://doi.org/10.1080/08824096.2011.565280

Hayes, A. F. (2013). *Methodology in the social sciences. Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford.

Hovland, C. I., Janis, I. L., & Kelley, H. (1953). *Communication and persuasion*. Yale University Press.

Jia, C. (2020). Chinese automated journalism: A comparison between expectations and perceived quality. *International Journal of Communication*, *14*(2020), 1–21.

Jia, C., & Gwizdka, J. (2020). An eye-tracking study of differences in reading between automated and human-written news. In F. D. Davis, R. Riedl, J. v. Brocke, P.-M. Léger, A. B. Randolph, & T. Fischer (Eds.), *Information systems and neuroscience* (pp. 100–110). NeuroIS Retreat. https://doi.org/10.1007/978-3-030-60073-0_12

Jia, C., & Johnson, T. (2021). Source credibility matters: Does automated journalism inspire selective exposure? *International Journal of Communication*, *15*(2021), 3760–3781.

Johnson, T. J., & Kaye, B. K. (2013). The dark side of the boon? Credibility, selective exposure, and the proliferation of online sources of political information. *Computers in Human Behavior*, *29*(4), 1862–1871. https://doi.org/10.1016/j.chb.2013.02.011

Jung, J., Song, H., Kim, Y., Im, H., & Oh, S. (2017). Intrusion of software robots into journalism: The public's and journalists' perceptions of news written by algorithms and human journalists. *Computers in Human Behavior*, *71*, 291–298. https://doi.org/10.1016/j.chb.2017.02.022

Knobloch-Westerwick, S., Mothes, C., & Polavin, N. (2017). Confirmation bias, ingroup bias, and negativity bias in selective exposure to political information. *Communication Research*, *47*(1). https://doi.org/10.1177/0093650217719596

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Liu, B., & Wei, L. (2019). Machine authorship in situ: Effect of news organization and news genre on news credibility. *Digital Journalism*, *7*(5), 635–657. https://doi.org/10.1080/21670811.2018.1510740

Liu, R., Jia, C., & Vosoughi, S. (2021). A transformer-based framework for flipping political polarity of news articles. *Proceedings of the ACM on Human–Computer Interaction*, *5*(CSCW1), 1–26. https://doi.org/10.1145/3449139

Liu, R., Wang, L., Jia, C., & Vosoughi, S. (2021). Political depolarization of news articles using attribute–aware word embeddings. *Proceedings of the 15th International AAAI Conference on Web and Social Media*, *15*(1), 385–396. https://ojs.aaai.org/index.php/ICWSM/article/view/18069

Metzger, M. J., & Flanagin, A. J. (2015). Psychological approaches to credibility assessment online. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (pp. 445–466). John Wiley & Sons. https://doi.org/10.1002/9781118426456.ch20

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. (2003). Credibility in the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. In P. Kalbfeisch (Ed.), *Communication yearbook* (Vol. 27, pp. 293–335). Erlbaum.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication*, *60*, 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Newhagen, J., & Nass, C. (1989). Differential criteria for evaluating credibility of newspapers and TV news. *Journalism Quarterly*, *66*(2), 277–284. https://doi.org/10.1177/107769908906600202

Perloff, R. M. (2015). A three-decade retrospective on the hostile media effect. *Mass Communication and Society*, *18*(6), 701–729. https://doi.org/10.1080/15205436.2015.1051234

Pettigrew, T. F. (2017). Social psychological perspectives on Trump supporters. *Journal of Social and Political Psychology*, *5*(1), 107–116. https://doi.org/10.5964/jspp.v5i1.750

Pew Research. (2014). *Political polarization in the American public*. https://www.pewresearch.org/politics/2014/06/12/appendix-a-the-ideological-consistency-scale

Shin, D. (2020). User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565. https://doi.org/10.1080/08838151.2020.1843357

Simons, H. W., Berkowitz, N. N., & Moyer, J. R. (1970). Similarity, credibility, and attitude change: A review and a theory. *Psychological Bulletin*, *73*, 1–16. https://doi.org/10.1037/h0028429

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology*, *89*, 845–851. https://doi.org/10.1037/0022---3514.89.6.845

Tandoc, E. C., Jr., Yao, L. J., & Wu, S. (2020). Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism*, *8*(4), 548–562. https://doi.org/10.1080/21670811.2020.1762102

Thurman, N., Dörr, K., & Kunert, J. (2017). When reporters get hands-on with robo-editing: Professionals consider automated journalism's capabilities and consequences. *Digital Journalism*, *5*(10), 1240–1259. https://doi.org/10.1080/21670811.2017.1289819

Vallone, R. P., Ross, L., & Lepper, M. R. (1985). The hostile media phenomenon: Biased perception and perceptions of media bias in coverage of the Beirut massacre. *Journal of Personality and Social Psychology*, *49*(3), 577–585. https://doi.org/10.1037/0022-3514.49.3.577

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society*, *22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Waddell, T. F. (2018). A robot wrote this? How perceived machine authorship affects news credibility. *Digital Journalism*, *6*(2), 236–255. https://doi.org/10.1080/21670811.2017.1384319

Waddell, T. F. (2019). Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly*, *96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Wang, S. (2021). Moderating uncivil user comments by humans or mchines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital Journalism*, *9*(1), 64–83. https://doi.org/10.1080/21670811.2020.1851279

Wojcieszak, M. (2019). What predicts selective exposure online: Testing political attitudes, credibility, and

social identity. *Communication Research*, *48*(5), 687–716. https://doi.org/10.1177/0093650219844868

Wu, Y. (2020). Is automated journalistic writing less biased? An experimental test of auto-written and human-written news stories. *Journalism Practice*, *14*(8), 1008–1028. https://doi.org/10.1080/

17512786.2019.1682940

Yamamoto, M., Lee, T. T., & Ran, W. (2016). Media trust in a community context: A multilevel analysis of individual- and prefecture-level sources of media trust in Japan. *Communication Research*, *43*(1), 131–154. https://doi.org/10.1177/0093650214565894

**About the Authors**

**Chenyan Jia** is a PhD candidate and Harrington dissertation fellow in the Moody College of Communication at The University of Texas at Austin. Her research interests include algorithmic bias, human-computer interaction, and computational social science. She is especially interested in how new media technologies such as automated journalism and misinformation detection algorithms are impacting media organizations and the public.

**Ruibo Liu** is a PhD student in the Department of Computer Science at Dartmouth College. His research interests include natural language processing, human-centered AI, and social computing. He studies political polarity detection, generative models for mitigating bias, and better evaluation metrics that reflect human preference.