Original Article

Does Polarizing News Become Less Polarizing When Written by an AI?

Investigating the Perceived Credibility of News Attributed to a Machine in the Light of the Confirmation Bias

Magdalena Wischnewski[1] and Nicole Krämer[1,2]

[1]Research Center for Trustworthy Data Science and Security, UARuhr, TU Dortmund, Germany
[2]Social Psychology: Media and Communication, University of Duisburg-Essen, Germany

Abstract: In this study, we examine how readers perceive the credibility of polarizing news purportedly written by a machine. In particular, we study whether a machine attribution can decrease the polarization inflicted by the self-confirmation bias. To that end, we expect that attitude-confirming polarizing news is perceived as less credible when attributed to a machine than when attributed to a human author. We assume this is due to the lower source credibility of machines and less emotional involvement. In a preregistered online experiment, we presented N = 508 participants with a polarizing news article attributed either to a human author or a machine. The article also either confirmed or disconfirmed participants' attitudes towards the polarizing issue. Our results show that participants did not differentiate between human and machine-attributed news. Moreover, we found no evidence that machine-attributed news affected the self-confirmation bias. However, we found that, while machine authors were perceived equally competent as human authors, they were perceived as less trustworthy. In addition, we found that the machine attribution induced less emotional involvement in terms of experienced enthusiasm but not experienced anger.

Keywords: polarizing news, automated journalism, credibility perceptions, self-confirmation bias, trustworthy automation

Artificial intelligence (AI) has long entered newsrooms. From the automation of back-end tasks, transcriptions, copy-editing, and content creation to news recommendation and news gathering, AI technologies accomplish various tasks. In particular, the commercial launch of large-language models (LLMs) such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018), GPT-4 (Generative Pre-trained Transformer 4; OpenAI, 2023), and Llama 2 (Touvron et al., 2023) allows for the automated creation of textual content with little to no human interference in the writing process. Based on sophisticated algorithms, these systems are more and more adopted by news organizations with the effect "that specific roles within the creation and distribution of news that once were associated with people are now being performed by software" (Guzman, 2019, p. 3), leading to terms such as automated journalism (Wölker & Powell, 2018), robot journalism (Clerwall, 2014), and algorithmic journalism (Graefe, 2016).

Likewise, tech companies are approaching news organizations to include journalistic content in their queries. As such big publishing houses such as, for example, the Associated Press (https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a) and Axel Spinger (https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism), have reported joining forces with OpenAI, allowing chat interfaces such as ChatGPT to include attribution and links to journalistic content directly.

With the introduction of these communicating machines, the question arises of how recipients perceive this communication and whether we can observe differences in human–human communication. Examining this question, previous research indicates that, while individuals can hardly distinguish machine-written from human-written texts (Clerwall, 2014; Graefe et al., 2018; Jung et al., 2017), differences arise when publishers disclose the author

attribution (i.e., algorithm versus human). Results suggest that knowing that a text is (purportedly) written by a machine affects how individuals perceive, for example, the text's credibility, albeit the direction of this effect is less clear. While some studies found that machine-generated news is perceived as more credible (Graefe et al., 2018; Jung et al., 2017; Liu & Wei, 2018; 2019; Wu, 2020), others found results in the opposite direction (Waddell, 2018; 2019). Nevertheless, some studies suggest that readers perceive machine-attributed content as similar to human-attributed content (Tandoc et al., 2020; Wölker & Powell, 2021; van der Kaa & Krahmer, 2014).

Moreover, when news is attributed to a machine, the possible effects on credibility are significant in the context of misinformation. Similar to previous discussions on the effects of social bots in spreading misinformation (see, e.g., Shao et al., 2018), malicious actors can use LLMs to create misleading content on a big scale. Likewise, accurate information about a disputed issue, such as vaccinations, might be even less accepted when written by a machine. Moreover, empirical evidence is needed to enforce legal requirements to ensure the disclosure of machine authorship.

In this context, with this paper, we aim to gather evidence about previously found differences between machine- and human-text attributions by investigating the effects of machine authorship on polarizing news content. In line with previous work, we found no difference between human- and machine-attributed content. Moreover, we found no evidence that a machine attribution would affect the occurrence of the self-confirmation bias. Nevertheless, we found differences between human and machine attributions regarding trustworthiness. Our participants perceived human-attributed text as more trustworthy than machine-attributed text, but both attributions were perceived as equally competent. In addition, we found that machine attribution induced lower levels of emotional involvement in our participants.

Literature Review

Automation in the Newsroom and its Effects on the Credibility Perceptions of News

With the emergence of automated journalism, questions arise about how recipients perceive these communicating machines in terms of credibility and whether differences in human–human communication can be observed. To answer this, in the following, we examine both existing media and communication theories as well as empirical findings that paint a complex picture of contradictory results.

Most prominently, examining possible differences between human and machine communication, Nass et al. (1994) propose in their widely used Computers are Social Actors (CASA) framework that individuals react to machines like humans. According to CASA, readers would perceive automated journalism as no different from traditional journalism. Indeed, empirical investigations support this notion (Tandoc et al., 2020; Wölker & Powell, 2021; van der Kaa & Krahmer, 2014).

However, theoretical predictions based on Sundar's (2008) Modality-Agency-Interactivity-Navigability (MAIN) framework and recent insights on *algorithm appreciation* (Logg et al., 2019) indicate that individuals favor machine-generated output over human-generated output. Formulated in the *machine heuristic*, Sundar (2008) argues in the MAIN model that individuals commonly apply epistemic assumptions associated with machines, such as being ideologically unbiased and objective, leading to the preference of machines over humans. Concepts such as "epistemic authority" (Carlson, 2019), "calculative objectivity" (Beer, 2017), and "mechanical objectivity" (Daston & Galison, 2010) echo these claims of "technologically inflicted promise of mechanical neutrality" (Gillespie, 2014, p. 181).

Going beyond theory, empirical support for a preference for machines over human authors comes from Sundar, Shyam and Kim (2019), who found that individuals were likelier to disclose sensitive information to a machine than another human. Similarly, insights by Logg and colleagues (2019) show that individuals adhered more to advice when the advice purportedly came from an algorithm than from a person, which led Logg et al. (2019) to coin the term algorithm appreciation. Moreover, Logg et al. (2019) suggest that "individuals may feel more comfortable with algorithmic advice in domains that feature a concrete, external standard of accuracy" (p. 91), which implies that algorithms are commonly perceived as objective, less ideologically biased, and, hence, more accurate, echoing the conceptualization of the machine heuristic. Such an appreciation or preference for machines over humans can also be found in empirical results on automated journalism (Graefe et al., 2018; Jung et al., 2017; Liu & Wei, 2018, 2019; Wu, 2020).

In contrast to the machine heuristic and results on algorithm appreciation, previous empirical results on *algorithm aversion* (Dietvorst et al., 2015; Dietvorst et al., 2018) indicate that individuals prefer humans over machines, especially when they see a machine err. This aversion might be grounded in algorithmically powered machines being less transparent and opaque, which relates to the infamous black box. These assumptions find empirical support in previous work on automated journalism, which could show that individuals perceived human-generated content as more credible than machine-generated content (Waddell, 2018; 2019). Because previous empirical results have been mixed and theoretical assumptions result in conflicting predictions, we refrain from posing a directional hypothesis but instead pose the following research question:

Research Question 1 (RQ 1): Are news articles purportedly written by an algorithm perceived as less, more, or similarly credible than those purportedly written by a human author?

Automated Pro- and Counter-Attitudinal News: The Case of Confirmation Bias

As outlined in the previous section, there are empirical and theoretical reasons to assume that news readers will perceive machine-attributed news differently than human-attributed news. While the empirical investigations cited above investigate the effects of machine attribution on different news genres, such as sports or finance (Graefe et al., 2018; Jung et al., 2017; van der Kaa & Krahmer, 2014; Wölker & Powell, 2021; Wu, 2020) as well as weather, the stock market and science (Waddell, 2018), only a few studies investigated effects of machine attribution on polarizing or disputed content. Although Waddell (2019) presented participants with a disputed news story (Trump dispute with Khan; Trump rejection of Paris Climate Accord), he did not control for individuals' attitudes towards these topics.

However, accounting for individuals' attitudes towards the content of a text is crucial. Research on the self-confirmation bias could show that credibility perceptions are guided by how the content aligns with the reader's attitudes. News confirming one's attitudes is quickly deemed credible, whereas news disconfirming one's attitudes is quickly deemed not credible (Metzger et al., 2010). Theoretically, the self-confirmation bias is grounded in conceptualizations such as cognitive dissonance theory (Festinger, 1976), which suggests that attitude-disconfirming information is perceived as a threat to one's identity, eliciting identity defense mechanisms and motivated reasoning (Kunda, 1990). Consequences of the self-confirmation bias can be highly detrimental, making readers, for example, more open to misinformation (e.g., Ecker et al., 2014) or leading them to reject accurate news (Wischnewski & Krämer, 2020).

When considering the self-confirmation bias in the context of automated journalism, a key question arises: Would assigning authorship to a machine potentially increase, decrease, or have no impact on the bias? Following the different ways in which a machine attribution could affect credibility perceptions in general (see previous section), different outcomes are possible. However, generally, to expect a machine attribution to show an effect on the self-confirmation bias, the attribution would have to affect the processing of attitude-confirming information differently than the processing of attitude-disconfirming information.

Theoretically, we see two different arguments suggesting such differences in processing strategies. First, if we follow the assumptions of the machine heuristic, readers would perceive a machine as more objective and less biased by ideological beliefs than a human counterpart. We presume that, according to this, machine attribution mainly affects the perception of attitude-disconfirming information. In particular, we assume that disconfirming news from an agent that is perceived as more objective should induce less attitude threat, making individuals more open to this information.

Second, if a machine is perceived as less credible than a human counterpart, this might affect the perception of attitude confirming information. We presume that confirming news from an agent that is perceived as less credible should reduce the identity affirmation effect, making individuals less prone to accept it on the mere basis of attitude confirmation.

Initial empirical investigations that tested the effects of machine attribution in the context of polarizing news indicate that, while machine-generated content did not affect the perception of attitude-disconfirming news, it reduced the effects of attitude-confirming news, making these less credible as compared to attitude-confirming news attributed to a human journalist (Wischnewski & Krämer, 2022). Following previous results, we hypothesize that:

Hypothesis 1 (H1): Readers perceive attitude-confirming news attributed to a machine as less credible than attitude-confirming news attributed to a human author.

Underlying Psychological Processes: Effects of Perceived Source Credibility and Emotional Reactions

To further scrutinize our hypothesis, we revisit the literature on confirmation bias, aiming to examine possible underlying psychological mechanisms for the effect we propose. In doing so, we suggest that the effects of perceived source credibility and affective reactions could explain hypothesis 1.

Perceived source credibility. The perceived credibility of a source is one of the most important factors when judging message credibility (see, e.g., Pornpitakpan, 2004). If the source credibility is low, message credibility is likewise expected to be low (Appelman & Sundar, 2016). Hence, the reduced source credibility of a machine attribution could theoretically explain why attitude-confirming news attributed to a machine is less credible than attitude-confirming news attributed to a human.

Empirically, results by Jia and Johnson (2021) support this notion. Investigating the effects of machine attribution on selective exposure, they found that, in one of their six conditions (gun rights story), participants preferred attitude-confirming human-attributed news over machine-attributed news. The authors explain their results by referring to differences in perceived source credibility. Because

machines were perceived as less credible sources, the effect of self-confirmation was reduced. Previous studies support these results, indicating similar results for human and machine-attributed source credibility (Wölker & Powell, 2018).

While both Jia and Johnson (2021) and Wölker and Powell (2018) conceptualized source credibility as a one-dimensional construct, we understand source credibility as an at least two-dimensional construct with the dimensions of *perceived trustworthiness* and *perceived expertise* (Hovland et al., 1953; Pornpitakpan, 2004) (see Metzger et al., 2003 for an overview of source credibility dimensions).

Both sub-dimensions are theoretically highly relevant not only in the context of news credibility but also in the context of human-machine communication. To that end, from the perspective of news credibility, the perceived trustworthiness of the source is understood as "the communicator's motivation to tell the truth about a topic" (Metzger et al., 2003, p. 297). From the perspective of human-machine communication, perceived trustworthiness is understood as the machine's ability, benevolence, and integrity (Mayer et al., 1995) to undertake a task. For both perspectives, perceived expertise is the communicator's ability or performance.

Based on this evidence, we hypothesize that the reduced credibility of machine-attributed, attitude-confirming news can at least partly be explained by machines being perceived as less credible sources than humans. When dividing source credibility into its two sub-dimensions, we hypothesize:

> *Hypothesis 2 (H2)*: The effect of a machine attribution on the perceived credibility of attitude-confirming news can partly be explained by machines being perceived as (H2a) less trustworthy, and (H2b) lower in expertise than human authors.

Affective responses. In addition to the hypothesized effects of source credibility, we argue that the hypothesized effect in H1 might also be related to readers' affective responses. Previous research shows that the self-confirmation bias is not an entirely cognitive process but is accompanied by affective responses (Suhay & Erisen, 2018; Wischnewski & Krämer, 2021). Suhay and Erisen (2018) found that the experienced affective states of enthusiasm and anger mediated the effects of self-confirmation on the perceived argument quality of news. The authors argue that individuals, upon perceiving attitude confirming news, experience a positive valuation of their social identity. In other words, encountering attitude-confirming information made participants feel more enthusiastic, leading to higher ratings, whereas encountering attitude-disconfirming information made participants feel angrier, leading to lower ratings of argument strength.

Assuming that a machine attribution only affects the perceived credibility of attitude-confirming news (see H1), we speculate that a machine attribution reduces the positive emotion participants experience when encountering attitude-confirming news but not the negative emotion when encountering attitude-disconfirming news. More formally stated:

> *Hypothesis 3 (H3)*: The effect of machine attribution on the perceived credibility of attitude-confirming news can partly be explained by machines reducing the positive feeling (enthusiasm) accompanying attitude confirmation. We assume to find no effect for negative feelings (anger).

Method

The study received ethical approval from the ethics committee of the University of Duisburg-Essen. The authors have preregistered this research with an analysis plan, which is retrievable at https://osf.io/rkd6v/.

Sample, Experimental Design, and Procedure

We recruited 508 participants via the online survey platform Prolific. Of these, 258 identified as female, 243 as male, four as nonbinary, and three preferred not to say. Participants mean age was $M = 29.74$ ($SD = 9.71$) and ranged from 18 to 73 years. Forty-eight participants indicated they received at least a middle school degree, 123 at least a high school degree, 54 completed an apprenticeship, 267 at least a university degree (Bachelor or Master), and 16 indicated none of the above. Hence, participants were relatively young and relatively educated.

To test our hypothesis, we conducted a 2×2 between-subjects design with the independent factors of *author attribution* (machine vs. human) and *attitude confirmation* (confirmation vs. disconfirmation). To manipulate the author attribution, participants in the machine group read: "The following news story you are about to read is automatically generated by an algorithm named AlgoInfo. AlgoInfo automatically generates content without human supervision. To ensure the quality of the algorithm, it was programmed to follow the common qualitative standards". For the human author, the participants read: "The following news story you are about to read is written by a human staff reporter, Kim Haas. Kim Haas has been educated as a journalist following the common qualitative standards". Additionally, the author's name was repeated in the article's byline. In total, $n = 216$ participants read news attributed to a machine, and $n = 212$ news attributed to a human.

The factor attitude confirmation referred to the agreement between participants' attitudes and the attitude slant

of the news article. Participants' attitude was measured on a 7-point Likert scale (from 1 = *completely disagree* to 7 = *completely agree*), and the article slant was varied between support for or opposition to gender-neutral language (see more in the next section). To arrive at a grouping variable, we created a matching variable of participants' attitudes towards the selected news story with the article they were presented. As a result, n = 232 participants read a news article confirming their attitude, and n = 196 read news that was disconfirming. As a self-confirmation bias can only be expected if individuals hold a prior attitude for or against an issue, participants who did not express an attitude (for the measurement, see below) were excluded from further analysis (n = 80) (for similar procedures, see, e.g., Brenes-Peralta et al., 2021).

At the beginning of the study, we asked participants to provide standard demographic data. Subsequently, we introduced participants to the task and randomly assigned them to the experimental conditions where they were told they were about to read a news article by a human journalist or a news algorithm. After reading the article, participants evaluated the author's credibility, how they felt when reading the news, and how credible they found it. The study closed with the measurement of two control variables (attitude towards algorithms, and prior experience with algorithms) and a manipulation check, after which participants were fully debriefed.

Stimulus Material

We asked participants to read the news article about the polarizing topic of gender-neutral language in Germany. To confirm the polarizing nature of the selected topic, we conducted a pretest with N = 51.

Both articles represented supporting or contradicting arguments for gender-neutral language usage. The length of each article was held constant and was between 236 and 247 words. To ensure that possible differences in the perceived credibility were not a result of differences in argument quality, in both articles, text wording was kept as similar as possible using negations. A t-test confirmed that participants perceived both articles as similarly credible $t(506) = 0.32$, $p = .532$ ($M_{against} = 4.71$, $SD_{against} = 1.35$; $M_{pro} = 4.78$, $SD_{pro} = 1.39$).

Measures

Dependent Variable

We adopted Appelman and Sundar's (2016) message credibility scale to assess the perceived credibility of the news articles. The scale consists of three adjectives on which participants were asked to rate how well each adjective described the news article: accurate, authentic, and believable (from 1 = *describes very poorly* to 7 = *describes very well*). All three ratings were summarized into one mean score with Cronbach's α = .864.

Additional Measures

Because the two dimensions, perceived trustworthiness, and perceived expertise, largely comprised a source's credibility (Pornpitakpan, 2004), we assessed both constructs using the subscales of trustworthiness and competence of McCroskey and Teven's (1999) scale. Both constructs are measured via six items on 7-point semantic differential scales with a Cohen's α = .918 for perceived competence and a Cohen's α = .906 for perceived trustworthiness.

While measures such as the Positive and Negative Affect Schedule (Watson et al., 1988) would be suitable to assess affective states that result from nonpolarizing information, as theorized above, we were interested in participants' emotional responses related to a possible identity affirmation (confirmation condition) or identity threat (disconfirmation condition) after being exposed to a polarizing news article. To capture such identity-protective and identity-bolstering affective states, we opted for self-report measures following affective intelligence theory (Marcus et al., 2000), previously used to examine emotional reactions for confirmation biases (Suhay & Erisen, 2018). Following Marcus et al. (2006), we used semantic emotional markers to assess the latent concepts of enthusiasm and anger. To that end, participants were asked to rate on a 7-point Likert scale, ranging from 1 (= *strongly disagree*) to 7 (= *strongly agree*), their experience of enthusiasm (enthusiastic, hopeful, proud; Cohen's α = .918) and anger (angry, outraged, disgusted; Cohen's α = .904).

Because attitudes towards algorithms have previously been found to affect how individuals perceive text attributed to machines, we additionally controlled for participants' attitudes. To do so, we adapted items created by Waddell (2018). Measured on a 7-point Likert-type scale (from 1 = *strongly disagree* to 7 = *strongly agree*), participants indicated their agreement with the following statements: (1) if an algorithm does a job, then the task was done objectively, (2) if an algorithm does a job, then the work was error-free, (3) if an algorithm does a job, then the work was unbiased, and (4) if an algorithm does a job, then the task was done accurately. All ratings were summarized into one mean score with Cronbach's α = .797.

Manipulation Check

At the end of the study, we asked participants whether they could recall the author's attribution of the news article they read. Participants could choose from 1 = "*a machine*", 2 = "*a human journalist*", and 3 = "*I cannot recall*". A chi-square test with the independent variable attribution and the dependent variable attribution recall indicated that the

Table 1. Descriptive results by experimental condition of the dependent measure perceived source credibility and the additional measures perceived source credibility (trustworthiness & expertise), enthusiasm, and anger

Variable	Author attribution	Confirmation	*M*	*SD*
Perceived message credibility	machine	confirmation	5.21	1.30
	human	confirmation	5.28	1.05
	machine	disconfirmation	4.21	1.35
	human	disconfirmation	3.97	1.43
Perceived trustworthiness	machine	confirmation	5.15	0.96
	human	confirmation	5.37	0.86
	machine	disconfirmation	3.99	1.24
	human	disconfirmation	4.29	0.96
Perceived expertise	machine	confirmation	5.38	1.04
	human	confirmation	5.39	0.93
	machine	disconfirmation	4.24	1.29
	human	disconfirmation	4.35	1.13
Anger	machine	confirmation	1.58	0.98
	human	confirmation	1.61	1.07
	machine	disconfirmation	3.28	1.72
	human	disconfirmation	2.86	1.51
Enthusiasm	machine	confirmation	4.46	1.38
	human	confirmation	4.73	1.34
	machine	disconfirmation	2.63	1.52
	human	disconfirmation	2.63	1.31

manipulation was successful $\chi^2(2) = 395.88$, $p < .001$. In the machine attribution condition, 93.06% ($n = 201$) of the participants recalled the attribution correctly, and 96.23% ($n = 204$) of the participants in the human attribution condition recalled the human author correctly.

Results

To test RQ1 and H1, we ran two multiple linear regressions (see Table 1 for standardized coefficients, standard errors, and *p*-values of both models) with credibility perceptions as the dependent variable, the author attribution (model 1), the interaction of the author attribution and the congruency condition (model 2) as predictors. Both models also included the control variables age, gender, education, attitudes towards algorithms, and prior experience with algorithms. All continuous variables were mean-centered. The descriptive results of all measures by condition can be found in Table 1.

Research Question 1

In RQ1, we asked whether machine-attributed news is perceived as similarly credible to human-attributed news. We tested this in a linear regression (model 1), which was overall significant with $F(6, 421) = 7.53$, $p < .001$, but did not explain much of the variance of our dependent variable perceived message credibility (adj. $R^2 = .08$).

We found that the author attribution did not affect the perceived credibility of the news article (see Table 2 for coefficients), supporting previous results (see Tandoc et al., 2020; Wölker & Powell, 2019; van der Kaa & Krahmer, 2014). Interestingly, we also found that one of our control variables, attitudes toward algorithms, significantly predicted participants' credibility perceptions ($\beta = 0.32$, $SE = 0.08$, $p < .001$).

Hypothesis 1

First, we found strong support for a self-confirmation bias, with attitude-confirming articles being perceived as significantly more credible than attitude-disconfirming articles (see Table 2). Second, in H1, we hypothesized that the self-confirmation bias would play out differently, depending on the author's attribution. Although we found no direct effect of a machine attribution for RQ1, we did not want to rule out this possible interaction due to a suppressor effect. Concretely, we suggested that a machine attribution would decrease credibility perceptions for attitude-confirming news compared to a human attribution. We tested this in Model 2, which was overall significant with $F(8, 419) = 17.48$, $p < .001$, and which explains significantly more variance of our dependent variable perceived message credibility (adj. $R^2 = .23$).

Unlike hypothesized, we found no significant interaction effect between the author attribution and the confirmation condition (see Table 1 for coefficients and Figure 1 for a

Table 2. Results of the multiple linear regression with the outcome variable perceived message credibility.

	Model 1		Model 2	
	β	*SE*	β	*SE*
Attribution (human)	−0.06	0.13	0.11	0.16
Confirmation (disconfirming)			−0.91***	0.17
Attribution × Confirmation			−0.36	0.24

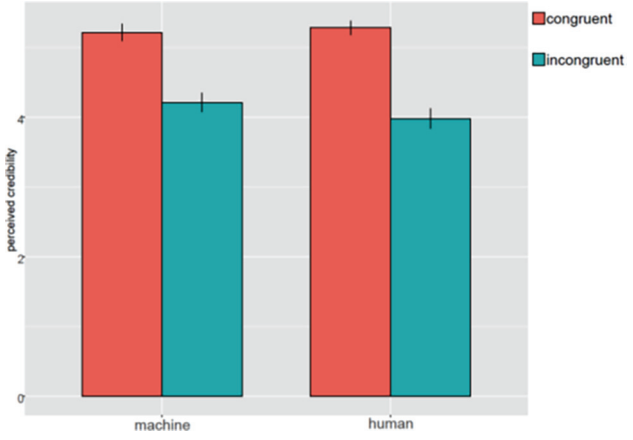

Figure 1. Perceived message credibility by experimental conditions.

visualization), disconfirming previous results by Wischnewski and Krämer (2022). These results indicate that machine attribution does not affect the occurrence of the self-confirmation bias.

Explorative Analyses of Hypothesis 1

Disclaimer: The following analyses and subsequent results were originally not part of our preregistration and were added post hoc. Although we measured participants' opinions on gender-neutral language via a 7-point Likert scale, the final congruency variable was binary, leading to a loss of scale level. To counter such a loss, we created a weighted congruency variable by combining participants' opinions on the polarizing issue with the information on the congruency condition. For this weighted congruency variable, the lower the variable's value, the stronger the disagreement with the participants' opinion, and the higher the value, the stronger the agreement with the participants' opinion. In the next step, we reran all analyses, which included the congruency variable.

Unlike the results reported above, reproducing the multiple linear regression with this weighted congruency variable to test whether a machine attribution would reduce the perceived credibility of attitude-confirming news supported H1. The regression was significant with $F(14, 413) = 12.39$, $p < .001$, adj. $R^2 = .27$). Inspecting the interaction term, we found that the interaction of the author attribution and the weighted congruency variable was significant (β = −0.13, *SE* = 0.05, $p = .019$). For machine attribution, the perceived credibility of news was reduced with increasing congruency. We assume this effect was masked when reducing the scale level to a binary congruency variable as reported in 4.2.

Hypothesis 2

In H2a and H2b, we suggested that the effect of machine attribution on the perceived credibility of attitude-confirming news could be explained by differences in perceived source credibility. Although we found no indication of such an interaction effect through the registered analysis (see section 4.2), we did not want to rule out possible underlying effects. Hence, to analyze our hypotheses, we ran a moderated mediation, using model 7 of PROCESS for R Version 4.0.1 with attitude confirmation as the independent variable, the variables perceived expertise and perceived trustworthiness as parallel mediators, author attribution as a moderator, and perceived credibility as the dependent variable. For the hypothesis testing, we used bootstrapping procedures, computing 1,000 bootstrapped samples with a confidence interval of 95%.

The bootstrapped results for the moderated mediation indicate neither a significant effect for perceived expertise (LLCI = −0.19, ULCI = 0.30) nor for perceived trustworthiness (LLCI = −0.11, ULCI = 0.19) (see Figure 2 for path coefficients). Hence, the moderated mediation in H2 was not supported. However, we found support for previous results, which indicate that the perceived credibility of the source mediates the effect of attitude confirmation on perceived message credibility.

Additionally, we found through the moderated mediation analysis that the author attribution significantly predicted levels of perceived trustworthiness (B = −0.30, *SE* = 0.14, $p = .036$), with the human attribution being perceived as more trustworthy than the algorithm attribution. In contrast, the attribution did not affect levels of perceived expertise (B = −0.11, *SE* = 0.16, $p = .489$). Taken together, these results suggest that, while both attributions were perceived as equally competent, they were not perceived as equally trustworthy. However, these differences in perceptions did not translate into differences in perceived message credibility.

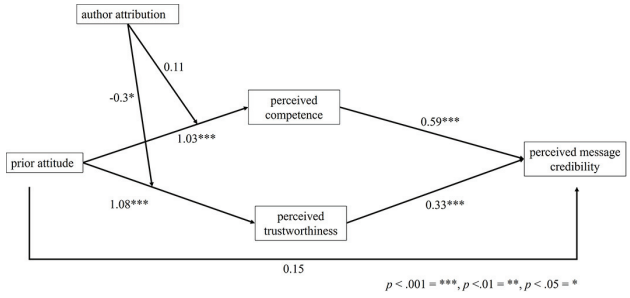

Figure 2. Path coefficients for the moderated mediation with the mediating variables perceived trustworthiness and perceived task expertise.

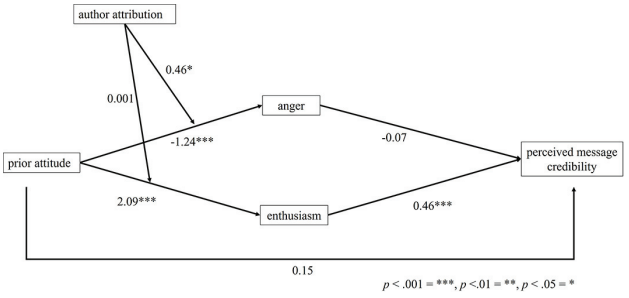

Figure 3. Path coefficients for the moderated mediation with the mediating variables enthusiasm and anger.

Rerunning the analysis with the exploratory variable *weighted congruency* to test H2 resulted in similar findings as described here. Results are reported in the supplementary material (see "Explorative Analyses" at https://osf.io/rkd6v/).

Hypothesis 3

In H3, we proposed that the effect of machine attribution on the perceived credibility of attitude-confirming news could be explained by reduced enthusiasm (but not anger). Similar to hypothesis 2, we analyzed this using the moderated mediation model 7 of PROCESS for R Version 4.0.1 with attitude confirmation as the independent variable, the variables experienced enthusiasm and experienced anger as parallel mediators, author attribution as the moderator, and perceived credibility as the dependent variable. For the hypothesis testing, we used bootstrapping procedures, computing 1,000 bootstrapped samples with a confidence interval of 95%.

First, we found support for previous findings (see Suhay & Erisen, 2018) on the confirmation bias, indicating that experienced enthusiasm mediated the effect of self-confirmation on perceived credibility. However, the bootstrapped results for the moderated mediation indicate neither a significant effect for the variable experienced anger (LLCI = −0.01, ULCI = 0.11) nor for experienced enthusiasm (LLCI = −0.38, ULCI = 0.12) (see Figure 3 for all path coefficients). Hence, the moderated mediation in H3 was not supported.

Additionally, as part of the moderated mediation analysis, we found that the author's attribution affected levels of experienced anger. Participants who read news attributed to a machine experienced lower levels of anger than participants who read news attributed to a human (B = 0.46, *SE* = 0.19, p = .025). However, the attribution did not affect levels of experienced enthusiasm.

Rerunning the analysis with the exploratory variable *weighted congruency* to test H3 resulted in similar findings as described here. Results are reported in the supplementary material (see "Explorative Analyses" at https://osf.io/rkd6v/).

General Discussion

In this study, we investigated the perceived message credibility of machine-attributed news. In particular, we were interested in how machine-attributed news is perceived in the context of the self-confirmation bias. To that end, we hypothesized that news attributed to a machine could reduce the effects of the self-confirmation bias by decreasing the credibility of attitude-confirming news. In addition, we were interested in possible underlying processes of this effect, related to source credibility and affective responses.

We tested our hypotheses through an online experiment in which we asked participants to read a news article that was either attributed to a machine or a human. In addition, we manipulated the article's slant, showing one group of participants' arguments for a polarized issue and the other group's arguments against the same issue to create the necessary condition for evoking self-confirmation bias.

First, our results indicate that our participants did not perceive news as more or less credible when they were attributed to a machine (RQ1), which supports the theoretical considerations by Nass et al. (1994) in their Computers are Social Actors framework as well as previous null results in the context of automated journalism (see Tandoc et al., 2020; Wölker & Powell, 2021; van der Kaa & Krahmer, 2014). However, alternative explanations to this finding suggest that the attribution manipulation was not strong enough to show an effect or that participants did not believe the claimed author attribution. Concerning the former, our results for the manipulation check indicate that participants did, in fact, remember the author attribution they were allotted to, which we interpret in a way that the manipulation was strong enough to be noticeable. However, we have no evidence that our participants believed the manipulation. Additionally, the mean ratings of perceived source

credibility were significantly above the scale midpoint for both the machine and human attribution conditions, indicating a possible ceiling effect. We speculate that the respective descriptions of the authors could have induced such a ceiling effect. In turn, a ceiling effect would have masked possible comparative differences between the two conditions, leading to the null result.

Second, we hypothesized that machine attribution would affect the occurrence of the self-confirmation bias. In the self-confirmation bias, individuals perceive attitude- confirming news as more credible than attitude-disconfirming news. Based on previous results by Wischnewski and Krämer (2022), we suggested that a machine attribution would reduce the effect of attitude-confirming news but not attitude-disconfirming news (H1). We also hypothesized that this would be due to (1) individuals perceiving machines as less credible sources (H2), and (2) machine attributions inducing less positive affective reactions (H3).

Testing these hypotheses, we found none confirmed. Machine attribution did not affect the confirmation bias; there was no connection between perceived source credibility and affective responses. We assume these null results are related to the null results of the research question, which implied that participants did not differentiate between a human and a machine attribution, contradicting Wischnewski and Krämer (2022), who found a main effect for the author attribution, indicating that machines were generally perceived as less credible than humans. Hence, it seems likely that, for cases where a machine attribution is perceived differently than a human attribution, we could also expect an effect in the confirmation bias — an interesting outlook for future studies.

Adding to these results, we found the hypothesized interaction of opinion congruency and author attribution through an explorative analysis, which included the strength of the confirmation and disconfirmation. Attitude-confirming news attributed to a machine was perceived less credible than attitude-confirming news attributed to a human author. We assume this effect was masked when reducing the scale level to a binary congruency variable. These results also suggest that a machine author can reduce the polarizing effect only for people with strong views about a topic.

While we found no evidence to support our hypotheses related to perceived source credibility and affective responses, our results, nevertheless, indicate that a machine attribution affected these variables. Dividing source credibility into two sub-dimensions, we found that machines were perceived as similarly competent compared to humans but less trustworthy. Although this difference was not reflected in the perception of message credibility, it supports previous results for algorithm aversion. It also indicates that this aversion is not necessarily related to the abilities and competence of machines.

Moreover, we found that a machine attribution resulted in lower levels of anger than a human attribution. This reflects previous results by Liu and Wei (2019), who found that machine authors induced less emotional involvement. However, our results are more fine-grained by differentiating between the different emotions of enthusiasm and anger, showing significant results only for anger but not enthusiasm. Future studies should examine emotional involvement more closely in different contexts and by differentiating emotions.

Interestingly, we found that our measure for attitudes towards algorithms was a positive predictor for perceived message credibility, independent of the author attribution. To better understand this finding, we revisited another study investigating attitudes toward algorithms that used a similar measure. As reported above, Sundar, Shyam, and Kim (2019) found that individuals were more inclined to entrust their credit card information to an algorithm than to another person, moderated by the belief in machine authority. Similar to our results, Sundar, Shyam, and Kim (2019) also report a significant main effect on attitudes toward algorithms (their variable was called machine heuristic). This lets us speculate that the positive attitudes towards machines (instead of towards humans) and the understanding that machines are more capable than humans might be associated with a generally higher trait propensity to trust.

Implications

Various implications can be derived based on our results. First, our results indicate that attributing news to a machine (here: a news algorithm) did not reduce the news's perceived credibility. As more and more news agencies employ such automated systems, our findings imply that this development would not harm readers' news reception. Moreover, unlike previous studies, we found no evidence in our data that a machine attribution would affect other credibility heuristics (here: the confirmation bias).

While we found no differences between a machine and a human attribution concerning the perceived message credibility, differences still existed regarding perceived trustworthiness. Based on our results, we cannot make strong claims about why we found this with the exception that we know it was not related to differences in perceived expertise. Yet, our results echo calls emphasizing the importance of establishing trust in automated systems (see e.g., Aoki, 2021; Glikson & Woolley, 2020; Kaplan et al., 2021).

Lastly, we want to point to the strong confirmation effect which we found. It seems that, while questions about the credibility of communicating machines are highly justified, any of such effects will inherently be overshadowed by

other, possibly much stronger, effects related to biased human reasoning.

Limitations and Future Directions

The presented study includes certain shortcomings that future studies should address. We discuss these limitations considering the methodological choices but also theoretical limitations. Concerning methodological choices, the selected topic of gender-neutral language use, while appropriate to induce the self-confirmation bias, reflects an opinion piece rather than a classical news article. It might be commonly assumed that news algorithms do not voice opinions or subjective views; this choice might have led to reduced trustworthiness perceptions of the algorithm attribution. One way to avoid this is to select topics that represent polarizing facts (e.g., anthropomorphic climate change in US samples) which we did not choose as a pretest indicated only low polarizing potential for our sample. Second, as we hinted in the general discussion, the topic of automated journalism might be of special interest in the context of misinformation. To examine this, future studies could manipulate the accuracy of news articles and ask participants to identify false news. Third, we chose to assess individuals' credibility perceptions, which we assessed via self-reports. While this is a standard procedure, including a measure closer to individuals' behavior such as news selection and sharing intentions would be interesting.

Reflecting theoretical limitations, we limited our examination to the interaction of a machine attribution with just one of many credibility heuristics. While the self-confirmation heuristic has been shown to be a rather strong credibility heuristic, possibly overriding other heuristics (Metzger et al., 2010), it is important to understand the consequences of algorithm attribution in the context of various heuristics. Such heuristics are, for example, the authority heuristic, the bandwagon heuristic, or the expert heuristic (Sundar, 2008).

Conclusion

In this work, we asked whether a machine attribution is perceived similarly credible as compared to a human attribution. In addition, we were interested in how a machine attribution would affect the common credibility heuristic, the self-confirmation bias, and which underlying processes could explain this. To that end, we additionally investigated perceived source credibility and affective responses. To answer these questions, we asked 508 participants to read either a purportedly machine or purported human-generated news article and asked them how credible they found the content. To induce the confirmation bias, we also varied whether the news article presented content for or against a polarizing issue.

Our results indicate that participants did not differentiate between a machine and a human attribution. Moreover, we found no evidence that a machine attribution would affect the occurrence of the self-confirmation bias. However, we found differences between the human and machine attribution in terms of trustworthiness, with humans being perceived as more trustworthy than machines but not as more competent. In addition, we found that a machine attribution induced lower levels of emotional involvement in our participants, particularly lower levels of enthusiasm due to attitude-confirming news. Our work contributes an empirical investigation to the extant literature on human-machine communication which is grounded in social and communication psychology.

References

Aoki, N. (2021). The importance of the assurance that "humans are still in the decision loop" for public trust in artificial intelligence: Evidence from an online experiment. *Computers in Human Behavior, 114*, Article 106572. https://doi.org/10.1016/j.chb.2020.106572

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly, 93*(1), 59–79. https://doi.org/10.1177/1077699015606057

Beer, D. (2017). The social power of algorithms. *Information Communication and Society, 20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Brenes-Peralta, C., Wojcieszak, M., & Lelkes, Y. (2021). Can I stick to my guns? Motivated reasoning and biased processing of balanced political information. *Communication & Society, 34*(2), 49–66. https://doi.org/10.15581/003.34.2.49-66

Carlson, M. (2019). News algorithms, photojournalism and the assumption of mechanical objectivity in journalism. *Digital Journalism, 7*(8), 1117–1133. https://doi.org/10.1080/21670811.2019.1601577

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice, 8*(5), 519–531. https://doi.org/10.1080/17512786.2014.883116

Daston, L., & Galison, P. (2010). *Objectivity*. Princeton University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science, 64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Drummond, C., & Fischhoff, B. (2019). Does "putting on your thinking cap" reduce myside bias in evaluation of scientific evidence? *Thinking and Reasoning, 25*(4), 477–505. https://doi.org/10.1080/13546783.2018.1548379

Ecker, U. K. H., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation.

Memory and Cognition, 42(2), 292–304. https://doi.org/10.3758/s13421-013-0358-x

Festinger, L. (1976). *A theory of cognitive dissonance*. Stanford University Press.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167–193). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals, 14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Graefe, A. (2016, January). *Guide to automated journalism*. Columbia Journalism Review. https://www.cjr.org/tow_center_reports/guide_to_automated_journalism.php#summary-outlook

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

Guzman, A. L. (2019). Prioritizing the audience's view of automation in journalism. *Digital Journalism, 7*(8), 1185–1190. https://doi.org/10.1080/21670811.2019.1681902

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

Jia, C., & Johnson, T. J. (2021). Source credibility matters: Does automated journalism inspire selective exposure? *International Journal of Communication, 15*, 3760–3781. https://ijoc.org/index.php/ijoc/article/view/16546

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

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in Artificial Intelligence: Meta-Analytic Findings. *Human Factors*. https://doi.org/10.1177/00187208211013988

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Liu, B., & Wei, L. (2018). Reading machine-written news: Effect of machine heuristic and novelty on hostile media perception. In M. Kurosu (Ed.), *Human-computer interaction. Theories, methods, and human issues* (Vol. 10901, pp. 307–324). Springer International Publishing. https://doi.org/10.1007/978-3-319-91238-7

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

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151*(April 2018), 190–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Marcus, G. E., MacKuen, M. B., Wolak, J., & Keele, L. (2006). The measure and mismeasure of emotion. In D. Redlawsk (Ed.), *Feeling politics: Emotion in political information processing* (pp. 31–45). Palgrave Macmillan US.

Marcus, G. E., Neuman, W. R., & MacKuen, M. B. (2000). *Affective intelligence and political judgement*. University of Chicago Press.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McCroskey, J. C., & Teven, J. J. (1999). Goodwill: A reexamination of the construct and its measurement. *Communication Monographs, 66*(1), 90–103. https://doi.org/10.1080/03637759909376464

Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & Mccann, R. M. (2003). Credibility for the 21st Century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Annals of the International Communication Association, 27*(1), 293–335. https://doi.org/10.1080/23808985.2003.11679029

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication, 60*(3), 413–439. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

OpenAI. (2023). *GPT-4 Technical report*. https://arxiv.org/abs/2303.08774

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K. C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications, 9*(1), 1–16. https://doi.org/10.1038/s41467-018-06930-7

Suhay, E., & Erisen, C. (2018). The role of anger in the biased assimilation of political information. *Political Psychology, 39*(4), 793–810. https://doi.org/10.1111/pops.12463

Sundar., Shyam, S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Conference on Human Factors in Computing Systems - Proceedings*, 1–9. https://doi.org/10.1145/3290605.3300768

Sundar, Shyam. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital Media, Youth, and Credibility*, 73–100. https://doi.org/10.1162/dmal.9780262562324.073

Tandoc, E. C., 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

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., & Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models*. https://arxiv.org/abs/2307.09288

van der Kaa, H., & Krahmer, E. (2014). Journalist versus news consumer: The perceived credibility of machine written news. *The Computation Journalism Conference New York*, 1–4. http://www.poynter.org/latest-news/regret-the-error/205816/5-%0A; https://pure.uvt.nl/portal/files/4314960/c

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 and Mass Communication Quarterly, 96*(1), 82–100. https://doi.org/10.1177/1077699018815891

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication, 65*(4), 699–719. https://doi.org/10.1111/jcom.12164

Wischnewski, M., & Krämer, N. (2020). I reason who I am? Identity salience manipulation to reduce motivated reasoning in news consumption. In A. Gruzd, P. Mai, R. Recuero, Á. Hernández-García, C. S. Lee, J. Cook, J. Hodson, B. McEwan, & J. Hopke (Eds.), *Proceedings of the 11th International Conference on Social Media and Society* (pp. 148–155), New York, NY, USA.

Wischnewski, M., & Krämer, N. (2021). The role of emotions and identity-protection cognition when processing (mis) information. *Technology, Mind, and Behavior, 2*(1). https://doi.org/10.1037/tmb0000029

Wischnewski, M., & Krämer, N. (2022). Can AI reduce motivated reasoning in news consumption? Investigating the role of attitudes towards AI and prior-opinion in shaping trust perceptions of news. *HHAI2022: Augmenting Human Intellect*, 184–198. https://doi.org/10.3233/faia220198

Wischnewski, M., & Krämer, N. (2023, December 14). *Reducing motivated trust perceptions of news through AI?* [Data, Materials]. https://doi.org/10.17605/OSF.IO/RKD6V

Wojcieszak, M., Thakur, A., Ferreira Gonçalves, J. F., Casas, A., Menchen-Trevino, E., & Boon, M. (2021). Can AI Enhance People's Support for Online Moderation and Their Openness to Dissimilar Political Views? *Journal of Computer-Mediated Communication, 26*(4), 223–243. https://doi.org/10.1093/jcmc/zmab006

Wölker, A., & Powell, T. E. (2018). Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *journalism, 22*(1), 86–103. https://doi.org/10.1177/1464884918757072

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

History
Received September 12, 2023
Revision received May 5, 2024
Accepted May 12, 2024
Published online August 2, 2024

Authorship
Magdalena Wischnewski: Conceptualization, Formal analysis, investigation, Writing – original draft, Writing – review & editing; Nicole Krämer: Funding Acquisition, Writing – review & editing.

Conflict of Interest
We have no known conflict of interest to disclose.

Open Science
The data collected for the study in this paper, as well as the code to analyze the data, are publicly available via https://osf.io/rkd6v/ (Wischnewski & Krämer, 2023).

Funding
This work has been supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the https://uaruhr.de.
Open access publication enabled by Technical University of Dortmund; TU Dortmund.

ORCID
Magdalena Wischnewski
https://orcid.org/0000-0001-6377-0940
Nicole Krämer
https://orcid.org/0000-0001-7535-870X

Magdalena Wischnewski
Research Center for Trustworthy Data Science and Security
TU Dortmund
Joseph-von-Fraunhofer-Straße 25
44227 Dortmund
Germany
magdalena.wischnewski@tu-dortmund.de

Magdalena Wischnewski (PhD) is a PostDoc at the Research Center for Trustworthy Data Science and Security in Dortmund, Germany. She received her in Social Psychology in 2021 from the University of Duisburg-Essen. In her research, she investigates trust in autonomous systems, including trust calibrations, trust assessments, and the perception of these systems.

Nicole Krämer (PhD) is a Professor of Social Psychology: Media and Communication at the University of Duisburg-Essen. Her research interests include human-computer interaction and computer-mediated communication, especially social media. More specifically her research focuses on forms and effects of social media usage, related to impression management, self-disclosure, or social comparison. She also analyses the social effects of virtual agents and robots.