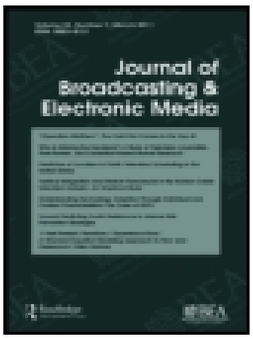

Journal of Broadcasting & Electronic Media

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/hbem20

The Nudging Effect of Accuracy Alerts for Combating the Diffusion of Misinformation: Algorithmic News Sources, Trust in Algorithms, and Users' Discernment of Fake News

Donghee Shin, Kerk F. Kee & Emily Y. Shin

To cite this article: Donghee Shin, Kerk F. Kee & Emily Y. Shin (2023): The Nudging Effect of Accuracy Alerts for Combating the Diffusion of Misinformation: Algorithmic News Sources, Trust in Algorithms, and Users' Discernment of Fake News, Journal of Broadcasting & Electronic Media, DOI: 10.1080/08838151.2023.2175830

To link to this article: https://doi.org/10.1080/08838151.2023.2175830

JOURNAL OF BROADCASTING & ELECTRONIC MEDIA
https://doi.org/10.1080/08838151.2023.2175830

The Nudging Effect of Accuracy Alerts for Combating the Diffusion of Misinformation: Algorithmic News Sources, Trust in Algorithms, and Users' Discernment of Fake News

Donghee Shin[a], Kerk F. Kee[b], and Emily Y. Shin[c]

[a]College of Communication and Media, Zayed University, Dubai, United Arab Emirates; [b]College of Media and Communication, Texas Tech University, Broadway Lubbock, Texas, USA; [c]American Community School of Abu Dhabi, Abu Dhabi, United Arab Emirates

ABSTRACT

Research has revealed that the diffusion of misinformation online can be attributed to inaccuracies in human cognitive reasoning. We examine the nudging effect of Accuracy Alerts on users' ability to discern misinformation and how their trust moderates this effect. The results from a 2 (accuracy nudge: with nudge vs. no nudge) × 2 (news source type: algorithmic news vs. legacy non-algorithmic news) experiment showed significant main and interaction effects indicating that an algorithmic source effect is present in the process of nudge acceptance. Misinformation sharing intention was largely lower for legacy non-algorithmic news than algorithmic news, but there was a greater decrease in algorithmic news when nudging was employed.

The nudging principle has been accepted and used to directly reinforce and indirectly influence the behavior and decision-making of individuals and/or groups in numerous contexts (Sunstein & Thaler, 2014). Because of its effectiveness, nudging has been widely applied to user analytics, system design, and public policy arenas (Mertens et al., 2021). In the digital context, nudging has been conceptualized as an indirect form of interactive and psychometric design that influences users online behavior without limiting their freedom of choice. In the digital media literature, nudging has also been studied as a concept of affordances (e.g., media affordance), which allows users greater control over what, when, and how they view content online

CONTACT Donghee Shin donghee.shin@zu.ac.ae Zayed University, Abu Dhabi, United Arab Emirates
Supplemental data for this article can be accessed online at https://doi.org/10.1080/08838151.2023.2175830

(Vermeulen, 2022). In this paper, a nudge refers to an alert message that warns users of the potential inaccuracy of a news story online.

Recent research has considered nudging for media applications, such as the evaluation of media credibility (Bhuiyan et al., 2018) and the affordance of social interactions (Shin, 2022). Further, nudging has been suggested as a means to combat misinformation and fake news (Andı & Akesson, 2021), for example, in the form of fact-check alerts and flags. More specifically, a nudge can provide a flag to trigger a user's attention to and cognitive processing about the trustworthiness and reliability of a news story. Thus, content posted on news feeds can be accompanied by an alert flagging that the news source can factually be wrong (Bryanov & Vziatysheva, 2021). An Accuracy Alert can then be a kind of nudging device to trigger users' attention toward discerning online news accuracy (Pennycook et al., 2021). Given this, Accuracy Alerts (e.g., flagging tags, embedded tweets, and pop-up banners) can intervene the otherwise quick sharing by unknowing readers of online news that may be misinformation or fake news by nudging readers to take a moment to assess before sharing (Roozenbeek et al., 2021).

Against the backdrop of the rising problem of fake news, calls for evidence-based interventions aimed at mitigating the diffusion of misinformation have been increasing (Nekmat, 2020). Relevant studies have addressed how simple interventions via a nudge can trigger users to consider the accuracy of online information before sharing, thus decreasing the spread of misinformation (Pennycook et al., 2021). Fazio (2020) noted that pausing to consider why a headline is true or false contributes to a reduced intention to share misinformation. A number of methods on how to combat misinformation have been examined, and studies have investigated why people share false or misinformation on social media (Bryanov & Vziatysheva, 2021; Park, Chung, et al., 2022). Therefore, it is important to examine what influences users' intention to share misinformation on social media and how nudging can intervene such an intention. We specifically focus on misinformation as our study focuses on unknowingly sharing of fake news, as opposed to disinformation, which will be intentionally sharing of fake news, which cannot be easily intervened by nudging and Accuracy Alerts.

As a form of indirect intervention, Accuracy Alerts have been examined and used to nudge readers to judge news truthfulness, which can improve their discernment of what to share on social media (Vermeulen, 2022). Fact-checkers such as Full Fact, PolitiFact, FeedReflect, and Associated Press warn readers about information that has been the subject of fact-checks. While the effectiveness of such nudges has not been empirically confirmed (Choi et al., 2022), some of them, such as fact-checkers, have been growing in popularity, becoming a major trend in social media and algorithmic platforms (Roozenbeek et al., 2021).

Despite their importance, the effects of nudges and the mechanisms behind such Accuracy Alerts on user behaviors, particularly in an algorithm context, have received limited empirical research attention. Thus, the exact nature of how nudges work and with what effect, how users respond to them, and what impact they have on user behavior in relation to misinformation and fake news remains inconclusive. While there is extensive research in the fact-checking literature on the utility and effectiveness of fact-checking practices, technologies, and policies, there is a dearth of research on users' perception of nudges and the ensuing effect on content sharing. Little is known about the nudging potential of Accuracy Alerts to prevent users from sharing questionable information on social media. Consequently, scholars (e.g., Shin et al., 2022) have called for more theoretical and practical research on the nudging effect of Accuracy Alerts in combating the diffusion of misinformation on social media.

We respond to this call for empirical studies to narrow the research gap by investigating (1) whether the presence of Accuracy Alerts flagging fake news will lead users to refrain from sharing questionable news on social media and (2) how nudging effect is moderated by source type and trust (algorithmic source effect) in a Korean context. A majority of news consumers in Korea (89% of the population) search for news searches via online sources, including algorithmic media platforms (Shin, 2022). Therefore, it is critical to understand the diffusion of misinformation on social media and the intervention of such a phenomenon. Misinformation that spreads online have negative ramifications at the individual and societal levels as it affects political processes and outcomes. The nudging effect of Accuracy Alerts and the potential for news source type to moderate the nudging effect conditional upon users' trust in algorithmic media are also examined. Furthermore, the paper expands on the theoretical foundations of nudging using source effects and trust evaluation perspectives. Our attempts also offer practical guidelines on the implementation and use of Accuracy Alerts in the prevalent algorithmic news environment on social media.

Literature Review

The online diffusion of misinformation has become a serious issue on social media and in our society (Wang et al., 2019). Relevant literature has examined the use of nudges in helping trigger fact-checking by news consumers online (Epstein et al., 2021). Recent empirical research has also found that shifting user attention to accuracy improves the relevance of the news they share on social media (Pennycook & Rand, 2022). Given these, we explore users' news accuracy discernment in an algorithm context by testing the moderating roles of news source type and users' trust in news. As previous studies are primarily based on U.S. samples (e.g., Pennycook et al., 2020), we

extend the relationships to a Korean context. News consumption and trust in media are different in the two countries due to cultural values. Therefore, it is worthwhile to examine how misinformation and user behaviors may interact differently in a new culture. Following this, we can gain insights into how to effectively discourage users from sharing misinformation.

News Recommender Systems and Misinformation

Misinformation, such as the unknowing sharing of fake news, is a current trend in the complex media landscape in our society. Misinformation is defined as unknowingly or accidentally sharing false information presented in a manner that is intentionally or unintentionally designed to mislead people. The terms misinformation and fake news are both referenced to describe the fact that the information is untrue and false. While misinformation refers to false or misleading information in general, the motivation behind the spreading of misinformation is either being simply unknowing or done without the intention to manipulate others (Lewandowsky & van der Linden, 2021). On the contrary, the term fake news is often understood as a form of misinformation and disinformation, typically with fabricated, false information presented in ways that are similar to legitimate news media content (Lazer et al., 2018). Given the complex ambiguity of intent to manipulate or not in the context of this study, the term misinformation is chosen to include the related trends.

Although the phenomenon of misinformation is not new, the prevalence of online and social media has compounded and amplified its detrimental impacts exponentially. Not only is misinformation more easily shared on social media today, the algorithms that favor engagement (e.g., views, likes, retweets) has also led to the unfortunate rise of "popular" content being mistaken as credible information by news consumers without prudent judgment regarding its accuracy (Shin, 2021). Given these, disinformation designed with manipulative intent could subsequently be shared when perceived by unknowing many to be true, thereby becoming misinformation. As a result, misinformation could put firms and brands in negative spotlight and cause longer-term harm to corporate reputation and consumer trust if no appropriate intervention is carried out (Jahng, 2021). To effectively combat misinformation and mitigate its detrimental effects, communication professionals need to understand the mechanism through which it could impact consumer responses, including what triggers a pause before quick sharing behaviors.

Interestingly, this nudging effect has not been empirically tested in an algorithm context. There have been extensive studies on the effects of labeling the credibility of news sources (e.g., Nassetta & Gross, 2020), but the nudging effect of Accuracy Alerts on AI-powered media have been scarce. Given the widespread use of AI in social media and news recommendations,

it is worthwhile to understand how nudging effect is differentially applied to algorithmic systems. In the era of algorithmically mediated environments, misinformation can be particularly powerful, especially through social media by mass users (Lewandowsky & van der Linden, 2021). Thus, individual user decisions to adopt and diffuse misinformation have reached new levels of importance as they can be a micro-level driver of the macro-level large-scale spread of false information that collectively put democratic decision-making at risk (Vosoughi et al., 2018).

The nudging potential of Accuracy Alerts as an intervention can be particularly relevant because recommender systems have been regarded as one of the major mechanisms behind misinformation diffusion in the online sphere (Shin et al., 2022). For example, news recommender systems (NRSs) have recently been the subject of intense criticism for contributing to echo chambers and information cocoons, which can influence the dynamics of social media news sharing and the ways in which misinformation originates, becomes legitimized, and propagates (Cardenal et al., 2019). Online news networks provide ubiquitous access to users, allowing them to surf through relevant news using algorithmic platforms. Therefore, it is a relevant and timely effort to examine the antecedents of the discernment of misinformation conceptually and empirically as well as more general human reasoning models of the nudging effect of Accuracy Alerts. A better understanding of the motivations and mechanisms of the susceptibility to fall prey to misinformation can lead us to reduce the adoption and diffusion of misinformation (Talwar et al., 2020).

Research Questions

Despite the increasing interest in NRS and Accuracy Alerts, research on how these nudges influence users to process misinformation via NRSs has remained at the level of discussion. Although relevant research confirms the accuracy inductions in subsequent discernment and sharing intention, the question remains as to how effective and valid they are in the context of AI-powered news media. The use of Accuracy Alerts has so far been championed, yet their actual feasibility and performance have yet to be fully examined, particularly in an algorithm context. Therefore, this study explores the conceptual development of Accuracy Alerts in online news. It investigates whether the priming and triggering of Accuracy Alerts for fake news will nudge users against sharing questionable news on social media. The effect of Accuracy Alerts and the mechanism of spreading misinformation have remained open questions for NRSs and journalism. Given the gap in the literature on misinformation and considering the importance of nudges in NRSs, we propose the following research questions:

RQ1: How do Accuracy Alerts trigger users' attention to misinformation sharing on social media? What are the nudging effects of Accuracy Alerts on users' discernment of misinformation and subsequent sharing behavior?
RQ2: How do news source types and users' trust in AI-powered media alter the nudging effect of such Accuracy Alerts? What are the effects of source type and trust on users' behaviors toward news consumption, particularly attitudes toward news sharing?

Hypothesis Development

Our experimental model was designed to test the nudging effect of Accuracy Alerts on users' attitudes toward misinformation discernment (Figure 1).

Accuracy Nudging Effect: Alerts Away from Falsehood

Online users often share inaccurate news, despite being aware that the news could be the product of unknowing misinformation and/or deliberate disinformation (Andı & Akesson, 2021). By sharing questionable news with their social networks (Jahng, 2021) and interacting with these networks on social media (Talwar et al., 2020), users may develop the status of influential sources or even opinion leaders (Dwyer & Martin, 2017), gaining benefits from such a status and improving their social capital (Shin et al., 2022). However, it is important for these users to correctly arrive at the belief that a news story is accurate and factual before proceeding to share it on social media. Accuracy Alerts have been found to be a useful tool for achieving such an outcome. Relevant research has supported this intervention by identifying a causal link between nudges and assurances in the context of algorithmic journalism (e.g., Fazio, 2020). Consequently, Accuracy Alerts can serve as a

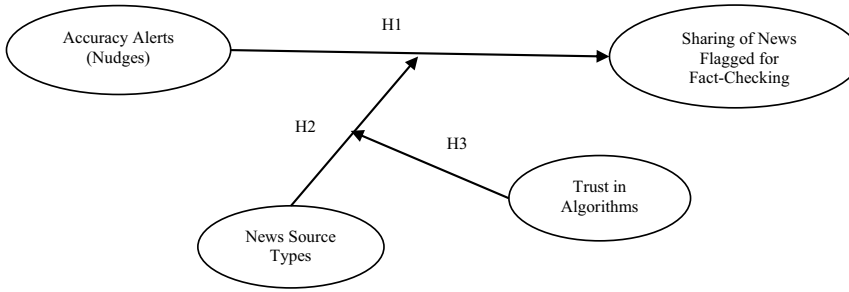

Figure 1. Hypothesized accuracy nudge model.

nudge, and thus, we present the following hypothesis:

H1: Accuracy Alerts nudge users toward reducing misinformation sharing intentions.

Algorithmic Source Effect: AI-Powered Media versus Legacy News Sources

News sources provide essential cues and intuitive heuristics that influence belief formation (Nekmat, 2020). Generally, the information from sources perceived as credible has a greater effect on users' beliefs, motivations, and attitudes than information associated with non-credible sources. Users trust news sources if they perceive the source as trustworthy, powerful, and if the sources are individuals, similar to themselves (Jia & Johnson, 2021). This kind of source judgment is applied to the algorithm context as users perceive the algorithmic source's credibility and believe the news it generates. Users are more likely to react to Accuracy Alerts for news sources that have been algorithmically personalized to them (Shin et al., 2022). In the algorithmic news environment, users understand that recommended news is selected, curated, and filtered by AI. When news is driven by algorithms, there is a tendency for users to believe that the recommended news is correct, accurate, and relevant to them. Therefore, Accuracy Alerts will more likely work when presented for algorithmic news because users are aware of the operation of algorithms in the background. Thus, we propose the following hypothesis.

H2: When Accuracy Alerts are absent, news from legacy media is more likely to be shared than news from an algorithmic NRS. Conversely, when Accuracy Alerts are present, news from algorithmic NRS is less likely to be shared than news from legacy media sources.

Trust Effects

Trust has been found to play a significant role in news readership and an even more critical role in NRS as it establishes an interacting feedback loop between users and the recommending mechanisms (Park, Chung, et al., 2022; Shin et al., 2022). When users trust an NRS, they tend to look for news beyond their own personal interests (Baden & Springer, 2017). When users believe that news recommendations are customized to their interests, they consider the service to be beneficial and are willing to engage further with the NRS (Mattis et al., 2022). When alerted to the potential inaccuracy of a news story, users cognitively dissociate news content from the source and discern the accuracy of the news story by linking it to their existing schema, knowledge, and experience relating to the source (Metzger et al.,

2015). During the evaluation, any credibility violation can decrease the overall credibility judgment of the perceivably credible news related to a trusted news source. When they are skeptical of algorithmic processes, they will be more influenced by online legacy news. As users are expected to share news believed to be credible in their social network, we hypothesize as follows:

H3: Users' trust moderates the nudging effect of Accuracy Alerts such that users' discernment to share news flagged for fact-checking will be greater for news from non-algorithmic legacy online news sources than news from an algorithmic NRS.

Methods

In this study, an online experiment was conducted with a 2 × 2 (Accuracy Alert: with alert vs. without alert × news source: algorithmic vs. non-algorithmic news) between-subjects design, which produced a set of four conditions (Table 1).

Data Collection and Sample

An experimental design was carried out through an online survey of 400 respondents (Table 2). To check for reliability and validity, six additional research participants completed a pretest questionnaire. Recommendations and preliminary results from this pretest (i.e., shorten minimum completion time for multiple-choice questions) were incorporated before distributing the questionnaire on a large scale online. The recruited participants for the main experiment were randomly assigned to one of the four groups, with no statistically significant differences between the groups in terms of gender, education, prior knowledge about algorithms, or media consumption of news. The four groups did not differ in terms of age, gender, or prior experience. The majority of the participants were Korean (92%), the average age was 31.76 (SD = 9.87), and the gender ratio was almost equal (female: 52.1%; male: 47.9%). Most of the participants (61.0%) were college graduates or had college-level education. With regard to employment status, 120 respondents were students, and 110 were full time workers. Of the collected responses, 10 incomplete responses with missing information were omitted.

Table 1. 2 × 2 experimental design.

	Nudged With Alerts	Non-Nudged Without Alerts
Algorithmic	Group 1	Group 2
Non-Algorithmic	Group 3	Group 4

Descriptive statistics show that the data were distributed approximately normally.

Procedures

On the welcome page, the participants were provided information about the study structure, the research goal, the procedure, and anonymity assurance. Next, they were introduced to the detailed content of the online experiment. Participants indicated informed consent, after which they were randomly assigned to one of the four experimental conditions. All the participants completed a pretest questionnaire measuring their personal trust in the news presented, experience with NRS and online news, prior social media news-sharing experience, and trust in algorithms. They were then randomly assigned to each of the two news source stimuli (i.e., algorithmic and non-algorithmic/legacy). The participants saw more than thirty headlines/news on COVID-19 and related issues. Those exposed to an online legacy source were shown an interface of the online news page aggregated by Daum, which has one of the highest market shares in Korea and the highest daily news clicks and readership rates. It is considered the most reputable news portal in the nation and was an appropriate stimulus to represent an online legacy news source for the study. The participants assigned to the algorithmic sources were shown news recommended by an NRS, which was run by Naver. The participants were required to search for news stories about coronavirus, coronavirus vaccine, and COVID-19 spread.

Overall, participants spent an average of 15 minutes on the news page ($SD = 2.74$). A factual news story on a relatively timely and current topic was selected to determine their trust in algorithms during the experiment. Specifically, news stories on COVID-19 and related health information that contained misinformation or incorrect/false information in the news were presented. To manipulate a nudge, half of the participants in each of the news source groups were exposed to an Accuracy Alert that popped up next to the news story to alert participants of content veracity and news inaccuracy in the news story. The alert was akin to a banner used by social media

Table 2. Attributes of respondents per experimental group.

	Experimental Group			
	1	2	3	4
Mean/median age (SD)	21.31/20.01 (3.10)	22.89/21.32 (3.07)	21.49/20.43 (3.01)	23.83/22.83 (2.98)
Gender (female %)	51	50	49	51
Prior nudge experience (year)	2.4	3.1	1.7	0.9
College education %	47	51	43	49

when a piece of news or information is flagged as "dubious or misinformation" by a third-party fact-checker.

Stimuli

The participants selected one of the NRSs or used their preferred applications. After signing the consent form, they completed a pre-screening questionnaire that rated their existing trust in algorithmic and online legacy news. The stimuli material differed in personal usage behavior, individual characteristics, the way users interacted with the NRS, and specific news search preferences. This was due to the algorithms' personalization in curating news, interaction with users, and reliance on existing data and the current trends in news agendas. The respondents downloaded an app from among a number of news apps, read news recommended through the app. They concluded the experiment by completing a questionnaire measuring their attitude, behavior, and intention to share the news. To validate the reliability of the responses, a series of confirmation questions were included in the survey. Furthermore, data reduction was performed on the initial data set in relation to the reliability and validity of the responses.

Measures

Manipulation checks and multiple control were performed for the experimental manipulation and measurement of the result.

Misinformation Sharing. News sharing concerns individuals' intention to share the news story and/or relevant information related it to on social media. Adapting prior measures on news-spreading intention (e.g., Dwyer & Martin, 2017), the participants ranked on a 7-point Likert-type scale (1 = strongly disagree and 7 = strongly agree) how likely they were to (1) suggest the news; (2) share the news; and (3) share opinions together with the news with others on social media ($\alpha = .88$, $M = 3.01$, SD =.88).

Trust in Algorithms. This refers to whether people perceive algorithms as a more trustworthy processor of their data than operations by humans. It was measured in a similar manner as in previous research (Schmitt et al., 2021; Shin, 2022) by asking the participants to rank their agreement (7 = strongly agree and 1 = strongly disagree) with three items (scale adapted from Shin (2021): (1) I think the algorithm is largely reliable and accurate in generating news recommendations, (2) I trust the algorithm in terms of data collection, analysis, and reflection, and (3) news recommendation by algorithms can be trusted due to their bias-free filtering ($\alpha = .87$, $M = 3.55$, $SD = 1.16$).

Analyses

A GLM univariate analysis of variance (ANOVA) was conducted to compare the mean ratings of the dependent variables (misinformation-sharing intention) between the Accuracy Alert and non-alert conditions (independent variables). Furthermore, we analyzed the indirect effect of trust as a moderator of news sharing. A median split was used to categorize trust ($M_{median} = 3.20$) into low and high levels. Using Welch's *F*-test, multiple one-way ANOVAs determined the mean values of trust ($F(3, 396) = 638.49$ $p < .001$; $\eta^2 = .829$; Welch statistics: 696.86) and misinformation sharing ($F(3, 396) = 83.66$, $p < .001$; $\eta^2 = .388$; Welch statistics: 144.41) as statistically significant between groups. Moreover, a pairwise *t*-test was performed to determine the moderation effect of trust on Accuracy Alerts and sharing intention (Table 3).

For Hypothesis 1—the nudging effect of Accuracy Alerts on misinformation sharing intention – an ANOVA was conducted to compare the participants in the two conditions nudged with Accuracy Alerts vs. participants with no alerts. The test showed a significant main effect ($F(1, 398) = 8.094$, $p < .005$) that users nudged with Accuracy Alerts reported a lower intention to share misinformation than those without Accuracy Alerts (Groups 1 + 3 vs 2 + 4). Given this, it can be inferred that Accuracy Alerts clearly play a significant role in discouraging user intention to share misinformation. It can also be inferred that Accuracy Alerts, if used properly, can effectively combat the spread of misinformation.

Hypothesis 2 predicted a positive impact of the news source type (algorithmic vs. legacy) on misinformation sharing. A one-way ANOVA was performed to compare the effect of news source type on misinformation sharing (Groups 1 + 2 vs 3 + 4). A one-way ANOVA revealed a statistically significant difference in sharing between the two groups ($F(1, 398) = 218.41$, $p < .001$). Thus, H2 was supported as the participants who received news from NRS relied significantly more on the nudged alerts compared to those who received news from an online

Table 3. Experiment results.

Group	1	2	3	4
Misinformation Sharing (Cronbach's α = .88)	Algorithmic NRS		Legacy Online Media	
	M:4.77 SD:2.102	M:5.61 SD:1.153	M:2.470 SD:1.095	M:2.820 SD:2.006
Trust in Algorithms (Cronbach's α = .87)	M:5.860 SD:.841	M:5.620 SD:1.042	M:2.780 SD:.732	M:1.610 SD:.665
	Nudged	Non-nudge	Nudged	Non-nudge

legacy source. It can be inferred that users who received news from an algorithmic NRS responded significantly more to the nudges than users who received news from an online legacy source.

Hypothesis 3 posited the moderation effect of trust on the Accuracy Alerts. We conducted a moderation analysis using PROCESS Macro 3.1 (Model 1), in which misinformation sharing was the dependent variable, Accuracy Alert was the independent variable, and trust was the moderator. The moderation test showed a significant interaction effect between the news source types and trust in news sharing (Table 4, $p < .05$, CI: −19.43, −3.941). Results indicate that for news from algorithmic sources, the role of trust on sharing becomes significant.

The results of the further analysis with *t*-tests also support the moderation effects. Pairwise *t*-tests showed significant mean pairwise differences between all conditions. The pairwise comparisons between the algorithmic source, low trust (M Algorithmic source, Low trust = 5.23), and legacy source, low trust (M Non-algorithmic source, Low trust = 2.54), treatments ($p < .001$) and between the algorithmic source, high trust (M Algorithmic, High trust = 5.23), and legacy source, high trust (M Non-algorithmic, High trust = 3.12), treatments ($p < .001$) were significant. Similarly, the pairwise comparison between the Accuracy Alerts, high trust (M_{Nudge} High trust = 4.67), and the Accuracy Alerts, low trust (M Nudge, low trust = 2.54), treatments ($p < .001$) and between the non-nudged/no alerts, high trust (M Non-nudged, High trust = 4.67), and non-nudged, low trust (M Non-nudged, Low trust = 3.06), treatments ($p < .001$) were significant. Accuracy Alerts lowered the intentions of users to share misinformation and the nudging effect was greater on algorithmic NRSs than in the non-algorithmic legacy sources. Users' sharing intention was also moderated by their trust level in the source type.

Table 4. Moderation Test by Hayes' PROCESS macro.

	Misinformation Sharing			
IV	coefficient	Standard Error	T	*p*
Trust	49,7284	17,5923	3.234	,0050
Source Type	68,5393	19,4913	3,7482	,0004
Trust X Source Type	-15,3194	5,3292	-1,9482	,0003
		Rsq	0,5923	,0010
		F	18,0329	

Table 5. Review of Hypotheses

Hypotheses	Key Findings	Results
H1. The effect of accuracy alerts on sharing	Accuracy alerts led to lower intention to share misinformation on social media.	Supported $p < .005$
H2. The positive impact of the algorithmic source type on sharing	Users who received news from algorithmic sources relied significantly more on the suggestions from this nudge than users who received news from legacy sources. Algorithmic sources had a significant effect on users' accuracy discernments.	Supported $p < .001$
H3. The moderation effect of trust on accuracy alerts	A significant interaction effect existed between news source types and trust for news-sharing behavior. When users had high trust in algorithms, they tended to more positively respond to accuracy alerts and less likely to share misinformation. When users had low trust in algorithms, they tended to consider accuracy alerts more lightly and were more likely to share misinformation.	Supported $p < .01$

Discussion

The results indicate that Accuracy Alerts can effectively nudge users and trigger an increase in discernment that intervenes in misinformation sharing on social media (Table 4). This point corroborates previous findings (Pennycook & Rand, 2022; Roozenbeek et al., 2021; Nekmat, 2020) demonstrating how nudges as heuristic cues can alter user action in sharing behaviors. Furthermore, the finding extends those from previous studies by highlighting that nudges in the form of Accuracy Alerts and heuristic cues could effectively influence user action in algorithmic media. In line with a prior finding, user attention is key to addressing misinformation, and a form of priming can help (Nekmat, 2020), as prompting users with an Accuracy Alert could trigger users' information processing, reminding them of the possible risks of sharing misinformation on social media. This nudge processing is consonant with ongoing research showing users' tendency to rethink their intention to share fake news with their social networks when they are reminded of attention to veracity (Pennycook et al., 2020). Our results largely correspond to existing literature but go a step further by clarifying the moderating effects of such a nudging effect on the discernment of misinformation. The algorithmic types of news sources and the levels of trust will likely influence the effectiveness of the Accuracy Alerts. We posit that users more willingly accept nudges (or respond more positively to nudges) when it is suggested by algorithms (i.e., NRS) rather than by humans or non-algorithmic services. We hypothesize that this is explained by the belief that human-free nudges appear less judgmental and will, thus, induce a greater sense of trust. This result contributes to the relevant literature on fake news

and misinformation by clarifying how misinformation nudges can cognitively influence users' intention to share misinformation on social media.

Algorithmic Source Effect

Our results imply that the credibility of algorithmic sources has significant effects on users' accuracy discernment when primed to judge misinformation (Table 5). The main effect found for Accuracy Alerts is in line with previous findings (e.g., Nekmat, 2020), with users exercising greater caution toward sharing news from algorithmic sources on social media compared to news from online legacy sources. Our result goes beyond this finding by illustrating that users were generally unwilling to share news from algorithmic sources that were flagged as containing misinformation. Instead, this effect of users' intention to share news flagged as misinformation when nudged by an Accuracy Alert was found to be stronger for news from an algorithmic source. This means that users are more aware of the fact that algorithmically curated and personalized news can be more vulnerable to fake news and misinformation and, therefore, more easily accept Accuracy Alerts when they are nudged to check the veracity and accuracy of the news they receive. This result can be explained by the theoretical perspective of source credibility and media credibility (Jia & Johnson, 2021). A user's positive response to algorithmic nudges is based on the user' perceived source credibility of algorithms. As users know that the news content is personalized by algorithms in line with what they expect or want to read, they more readily accept Accuracy Alerts from algorithms than a legacy news source because they believe that nudges are part of the algorithmic recommendations (Bryanov et al., 2020). Users of algorithmic NRSs deem Accuracy Alerts as algorithmically curated and checked and, thus, tend to believe them (Shin, 2022).

Trust Moderation

The algorithmic source effect is related to trust (Park et al., 2022). It can be inferred that the more users trust algorithm's mechanism, the more they are likely to respond to nudges. This inference is supported by the identified interaction effect between trust and news source types, which provides important insights into the role of trust in misinformation and fake news. It is understandable that users consider both cues when forming accuracy judgments (Bak-Coleman et al., 2022). When users have trust in a news source type, they tend to assume content credibility from these sources (Shin, 2021). A user's decision to assess the credibility of news is related to the news source type that they trust, that is, a source type that plays a more significant role in algorithmic sources than legacy sources. This implies the importance of the trust factor in algorithms – when users believe in algorithmic

mechanisms, they follow algorithmic suggestions, including Accuracy Alerts. When users do not have trust or lack thereof in algorithms, they are less likely to accept Accuracy Alerts because they are not convinced of the quality of algorithmic recommendations. Thus, one can expect trust in algorithms to influence users' assessment and the sharing of questionable news that was the subject of a nudge. Our finding bridges misinformation research – which lacks empirical referents – with processes that link nudge and trust, particularly in the context of users' sharing of misinformation in an algorithmic environment.

Implications

The goal of this study was to better understand, through a focus on algorithms, the processes contributing to users' responses to misinformation. The results of our study are relevant from both theoretical and applied standpoints. Theoretically, the discussions in this study make a relevant contribution to the dynamics of nudges, misinformation, algorithmic source effect, and news sharing on social media. Understanding the diffusion of misinformation requires us to examine the factors influencing users' discernment of misinformation. Evaluating the effect of Accuracy Alerts on news sharing from a user perspective can narrow the gap between theory and practice as it transcends NRS-specific concepts and theories by focusing on users' sensemaking process of misinformation and news sharing. A large part of online misinformation sharing is likely attributable to users' lack of attention to accuracy when considering what to share (Pennycook et al., 2021).

When users are nudged by algorithms, their attention to assess news accuracy increases with the interaction effect of trust in algorithm. This algorithmic source effect is arguably a theoretical advancement as little is known about source effects in the context of algorithmic services. The results of our study are in line with previous research that underlined the function of nudging in combating misinformation sharing (Pennycook et al., 2020) and the causal mechanisms underlying misinformation sharing as a result of a lack of cognitive reflection (Roozenbeek et al., 2021). We advance this finding by illustrating that perceived source credibility appears to be associated with misinformation sharing on social media and that nudge credibility encourages users' discernment of misinformation. Users tend to believe nudges when they rely on algorithms for news recommendations, believing that the nudges are part of the algorithmic personalization and curation. This tendency increases when users already have trust in algorithms. These results provide support for the theoretical argument concerning media source effects in an algorithm context. The algorithmic source effect implies the importance of trust in and credibility of algorithms. It also supports the importance of trust in human interactions with algorithms

(Shin et al., 2022) and further clarifies how credibility can be enhanced and how trust is established between users and algorithms. This theoretical implication further suggests that the effect of Accuracy Alerts should be understood from a user's sensemaking process: How users perceive the contextual factors regarding nudges, such as source type and trust. The results of this study show that drawing attention to the importance of accuracy can improve sharing judgment. Consistent with previous findings (e.g., Pennycook & Rand, 2022), this result indicates that Accuracy Alerts in news consumption do not only nudge the way in which users perceive and consume news online but can also help users avoid unwittingly engaging in misinformation sharing. The cognitive processes of news-sharing discernment have not been conceptually well-defined or explicitly operationalized, particularly in the context of AI-powered news platforms. Our attempt is meaningful because it empirically tested Accuracy Alerts from a user's sensemaking process to examine the nudging effect from a human perspective, not from a provider or algorithmic perspective.

The study shows the significance of Accuracy Alerts by illustrating how a nudge or cue that redirects attention to accuracy decreases the sharing of false content. This linkage serves as a foundation for theoretical contributions by establishing the relationship between accuracy and news sharing. This relationship has also been touted a great deal, despite being conceptually established. It also raises the need for further research relating to habit formation and continuous behavior in a long-term context, especially given the lack of user research on these topics. As the experiment sessions were short and did not test longitudinal effects, the results did not have enough explanatory power to warrant longer-term effects. This could be a limitation of this study and could be pursued in future research.

Practically, the nudge model provides useful suggestions for policymakers and social media platforms on how to manage and utilize Accuracy Alerts (Park, Chung, et al., 2022). The results provide algorithmic platform providers with guidelines on effective nudging prompts to choose from and cycle through when engaging users to increase the quality of information sharing online.

Conclusion

How effective is combating misinformation with nudges? This study provided experimental evidence of the nudging effect of Accuracy Alerts on users' discernment of fake news and behavior toward misinformation sharing from different sources on social media. While a wide range of motivations for sharing misinformation, including fun, entertainment, sarcasm, civic engagement, education, self-expression, social endorsement, and socialization (Madrid-Morales et al., 2021; Park, 2021), this study finds that users'

failure to judge information accuracy when consuming news can be one of the major factors of misinformation diffusion on social media. Our experimental evidence shows that priming attention toward the principle of accuracy can improve sharing judgment. The findings provide theoretically driven explanations and a more nuanced understanding of the sensemaking process of user sharing of questionable news on social media. The potential of Accuracy Alerts for fact-checking to impact algorithmic news media more than non-algorithmic sources offers a heuristic research focus for future studies examining the algorithmic effect of nudges on the perceived utility, trust, and credibility of algorithmic news platforms.

Limitations and Future Research

The results of the study should be treated with caution as the methods and design of our study comes with several limitations. First, we used a single message, which limits the scope of the findings. The results were drawn from a single news stimulus on a topic that affects a small portion of the larger population (e.g., COVID) instead of broader and general issues that influence the overall society. The use of a single message limits the replicability/generalizability of the findings and may not ensure that any effect is not idiosyncratic to the specific news story used in the experiment. Future studies may use multiple cases and longitudinal perspectives to increase the replicability and generalizability of the results. A longitudinal study would offer a useful framework that any scholars interested in user behavior related to misinformation in an AI environment should find useful. This insight will foster ensuring thinking and provoke excitement among researchers as we attempt to test and advance theory and solve radicalization problems in algorithmic platforms.

Second, we did not include the influence of individual traits and users' existing thoughts on media credibility and misinformation. Previous studies have shown the significant effects of user traits (e.g., audiences' skepticism toward news sources) on the evaluation of misinformation and fake news. Users may undergo selective news consumption like selective viewing, selective exposure, and selective processing based on their existing schema when judging misinformation. The process inevitably results in users, particularly among those who are highly politicized or highly oriented toward certain ideologies, evaluating news as being biased against their position and in favor of those in ideologically opposed positions and consequently perceiving the news as being more prejudiced and low quality. Hence it is reasonable to expect that individual traits play important roles in misinformation sharing behaviors.

Third, as this study focused on the effects of single nudges on misinformation, it missed diverse aspects of user behaviors related to misinformation.

It is possible that users may share news, accurate or fake, for many different reasons, and with different platforms on different networks. People may share misinformation purely for fun or for jumping on the bandwagon when a fake news story is trending. Future research should consider the user factors to influence the sharing of misinformation in different contexts.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Donghee Shin (Ph.D. & M.A., Syracuse University) is a professor at the College of Communication and Media Sciences at Zayed University. His research interests include digital journalism, human-computer interaction, and algorithmic media. His recent research in computer/algorithm as media addresses: embodied cognition, the subjective value of information, interaction, and interactivity.

Kerk F. Kee (Ph.D., the University of Texas at Austin) is an associate professor at the College of Media & Communication at Texas Tech University. His research primarily investigates the development, adoption, implementation, and ultimate diffusion of big data technologies in scientific organizations. He also studies the dissemination of health information in cultural communities and the spread of pro-environmental attitudes in modern societies.

Emily Y. Shin is a researcher at the Human-Computer Interaction Lab at Zayed University, Dubai, United Arab Emirates. Her research interests are in the areas of emerging information technologies and their impact on usability, interface, and user experience. She has participated in numerous interdisciplinary projects on human-AI research.

References

Andı, S., & Akesson, J. (2021). Nudging away false news. *Digital Journalism*, *9*(1), 106–125. http://doi.org/10.1080/21670811.2020.1847674

Baden, C., & Springer, N. (2017). Conceptualizing viewpoint diversity in news discourse. *Journalism*, *18*(2), 176–194. http://doi.org/10.1177/1464884915605028

Bak-Coleman, B., Kennedy, I., & Wack, M., Beers, A., Schafer, J.S., Spiro, E.S., Starbird, K. & West, J.D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behavior*, *6*, 1372–1380. https://doi.org/10.1038/s41562-022-01388-6

Bhuiyan, M., Vick, T., Mitra, K., & Horning, M. (2018). FeedReflect: A tool for nudging users to assess news credibility on Twitter. *Proceedings of the ACM Conference on Computer Supported Cooperative Work Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 205–208. https://doi.org/10.1145/3272973.3274056

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news. *PLoS ONE*, *16*(6), e0253717. https://doi.org/10.1371/journal.pone.0253717

Bryanov, K., Watson, B., & Pingree, R., Santia, M. (2020). Effects of partisan personalization in news portal experiment. *Public Opinion Quarterly*, *84*(S1), 216–235.

Cardenal, S., Aguilar-Paredes, C., Cristancho, C., & Majó-Vázquez, S. (2019). Echo-chambers in online news consumption. *European Journal of Communication*, *34* (4), 360–376. https://doi.org/10.1177/0267323119844409

Choi, S., Kim, H., & Gim, M. (2022). Do not read the same news! In *Companion Proceedings of the Web Conference 2022*. April 25-29, 2022, Lyon, France. https://doi.org/10.1145/3487553.3524936

Dwyer, T., & Martin, F. (2017). Sharing news online. *Digital Journalism*, *5*(8), 1080–1100.

Epstein, Z., Berinsky, A., Cole, R., Gully, A., Pennycook, G., & Rand, D. (2021). Developing an accuracy-prompt toolkit to reduce COVID-19 misinformation online. *Harvard Kennedy School Misinformation Review Harvard Kennedy School Misinformation Review*, *2*(3), 1–12. https://doi.org/10.37016/mr-2020-71

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr2020-009

Jahng, M. (2021). Is fake news the new social media crisis? *International Journal of Strategic Communication*, *15*(1), 18–36. http://doi.org/10.1080/1553118X.2020.1848842

Jia, C., & Johnson, T. (2021). Source credibility matters. *International Journal of Communication*, *15*, 3760–3781. 19328036/20210005. 19328036/20210005

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Madrid-Morales, D., Wasserman, H., Gondwe, G., Ndlovu, K., Sikanku, E., Tully, M., Umejei, E. & Uzuegbunam, C. (2021). Motivations for sharing misinformation. *International Journal of Communication*, *15*, 1200–1219. 19328036/20210005. 19328036/20210005

Mattis, N., Masur, P., Möller, J., & van Atteveldt, W. (2022). Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommenderdesign. *New Media & Society*. http://doi.org/10.1177/14614448221104413

Mertens, S., Herberz, M., Hahnel, U., & Brosch, T. (2021). The effectiveness of nudging. *Psychological and Cognitive Sciences*, *119*(1), e2107346118. https://doi.org/10.1073/pnas.2107346118

Metzger, M., Flanagin, A., & Nekmat, E. (2015). Comparative optimism in online credibility evaluation among parents and children. *Journal of Broadcasting & Electronic Media*, *59*(3), 509–529.

Nassetta, J., & Gross, K. (2020). State media warning labels can counteract the effects of foreignmisinformation. *Harvard Kennedy School Misinformation Review*, *1*, 1–11. https://doi.org/10.37016/mr-2020-45

Nekmat, E. (2020). Nudge effect of fact-check alerts: Source influence and media skepticism on sharing of news misinformation in social media. *Social Media + Society*, *6*(1), 1–14. http://doi.org/10.1177/2056305119897322

Park, Y. J. (2021). *The future of digital surveillance: Why digital monitoring will never lose its appeal in a world of algorithm-driven AI*. University of Michigan Press.

Park, Y. J., Chung, J. E., & Kim, J. N. (2022). Social media, misinformation, and cultivation of informational mistrust: Cultivating Covid-19 mistrust. *Journalism*, 23(12), 2571–2590. https://doi.org/10.1177/14648849221085050

Park, Y. J., Lee, H., Jones-Jang, S. M., & Oh, Y. W. (2022). Digital assistants: Inequalities and social context of access, use, and perceptual understanding. *Poetics*, *101689*.

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

Pennycook, G., McPhetres, J., Zhang, Y., Lu, G., & Rand, G. (2020). Fighting COVID-19 misinformation on social media. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2022). Nudging social media toward accuracy. *The ANNALS of the American Academy of Political and Social ScienceThe ANNALS of the American Academy of Political and Social Science*, *700*(1), 152–164. https://doi.org/10.1177/00027162221092342

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Schmitt, A., Wambsganss, T., Soellner, M., & Janson, A. (2021). Towards a trust reliance paradox? *ICIS 2021 Proceedings ICIS 2021 ProceedingsIcis 2021 Proceedings*, *14*. https://aisel.aisnet.org/icis2021/ai_business/ai_business/14

Shin, D. (2022). Expanding the role of trust in the experience of algorithmic journalism: User sensemaking of algorithmic heuristics in Korean users. *Journalism Practice*, *16*(6), 1168–1191. https://doi.org/10.1080/17512786.2020.1841018

Shin, D. (2022). The perception of humanness in conversational journalism. *New Media & Society*, *24*(12), 2680–2704. http://doi.org/10.1177/1461444821993801

Shin, D., Lim, J., Ahmad, N., & Ibahrine, M. (2022). Understanding user sensemaking in fairness and transparency in algorithms. *AI & Society*. https://doi.org/10.1007/s00146-022-01525-9

Sunstein, R., & Thaler, H. (2014). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Talwar, S., Dhir, A., Singh, D., Virk, S., & Salo, J. (2020). Sharing of fake news on social media. *Journal of Retailing and Consumer Services*, *57*, 102197. http://doi.org/10.1016/j.jretconser.2020.102197

Vermeulen, J. (2022). To nudge or not to nudge: News Recommendation as a Tool to Achieve Online Media Pluralism. *Digital Journalism*, 10(10), 1671–1690. http://doi.org/10.1080/21670811.2022.2026796

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, *40*, 112552