NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges

MD MOMEN BHUIYAN, Virginia Tech, USA
MICHAEL HORNING, Virginia Tech, USA
SANG WON LEE, Virginia Tech, USA
TANUSHREE MITRA*, University of Washington, USA

Struggling to curb misinformation, social media platforms are experimenting with design interventions to enhance consumption of credible news on their platforms. Some of these interventions, such as the use of warning messages, are examples of *nudges*—a choice-preserving technique to steer behavior. Despite their application, we do not know whether nudges could steer people into making conscious news credibility judgments online and if they do, under what constraints. To answer, we combine nudge techniques with heuristic based information processing to design NudgeCred–a browser extension for Twitter. NudgeCred directs users' attention to two design cues: authority of a source and other users' collective opinion on a report by activating three design nudges—*Reliable, Questionable,* and *Unreliable*, each denoting particular levels of credibility for news tweets. In a controlled experiment, we found that NudgeCred significantly helped users (n=430) distinguish news tweets' credibility, unrestricted by three behavioral confounds—political ideology, political cynicism, and media skepticism. A five-day field deployment with twelve participants revealed that NudgeCred improved their recognition of news items and attention towards all of our nudges, particularly towards *Questionable.* Among other considerations, participants proposed that designers should incorporate heuristics that users' would trust. Our work informs nudge-based system design approaches for online media.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Empirical studies in HCI*; *User studies*;

Additional Key Words and Phrases: Social Media; Twitter; Nudge; News; Credibility; Misinformation; Fake News; Intervention; Heuristic; Bandwagon

ACM Reference Format:
Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 427 (October 2021), 30 pages. https://doi.org/10.1145/3479571

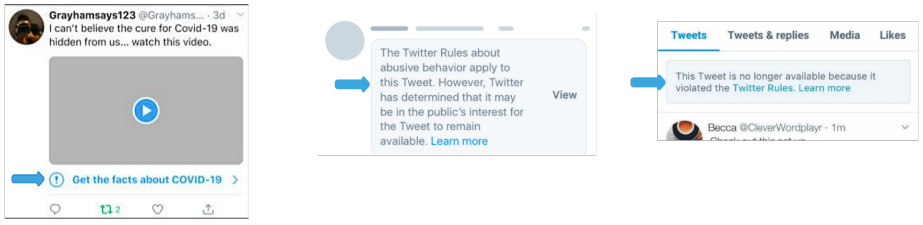
1 INTRODUCTION

Social media platforms have witnessed an unprecedented rise in misinformation around public issues (e.g., COVID-19 and the 2016 US Presidential Election [3, 137]). To tackle, they have responded by experimenting with various design interventions [104, 132]. These include attaching warning labels and links to show additional context from trusted sources (see figure 1 for an example). With

*A portion of this work was conducted while the author was at Virginia Tech.

Authors' addresses: Md Momen Bhuiyan, Virginia Tech, USA, momen@vt.edu; Michael Horning, Virginia Tech, USA, mhorning@vt.edu; Sang Won Lee, Virginia Tech, USA, sangwonlee@vt.edu; Tanushree Mitra, University of Washington, USA, tmitra@uw.edu.

2573-0142/2021/10-ART427 $15.00
https://doi.org/10.1145/3479571



(a) (b) (c)

Fig. 1. Three types of interventions (marked by blue arrows) currently employed by Twitter to tackle misinformation. Tweet (a) with a link to proper authority regarding COVID-19, (b) with a warning, and (c) removed. Here, both (a) and (b) are examples of nudges. Around the beginning of our work (July 2018), only (c) was operational. Twitter added others later.

an intent to raise awareness and lead users to investigate the veracity of a news item [106], such design interventions can act as *nudges*—a choice-preserving technique to steer behavior [121]. Nudges differ from methods such as real-time corrections that often act as mandates and can backfire [36]. Furthermore, nudges can overcome the scale issue that many real-time systems face who rely on limited number of expert fact-checkers and quality crowd workers [6]. Despite the benefits, there is little empirical evidence whether nudge-based socio-technical interventions affect users' perception of credibility of online information. Furthermore, what complicates such investigation is that people with strong ideological leaning may resist nudges [125]. Existing works, again, lack empirical evidence of the effects of ideological leaning on nudges regarding perception of credibility. Therefore, considering the constraints under which these nudges may or may not work is crucial. This paper does just that.

For the purpose of our investigation, we design nudges with heuristic cues, i.e., mental shortcuts that people often use to judge credibility [75]. The choice of heuristic cues over reflective ones reduces cognitive burden on users, given the immense amount of content users see online [33]. Incorporating design guides from nudge and heuristics literature [117, 128], we built NudgeCred which operationalizes three design nudges—*Reliable*, *Questionable* and *Unreliable*. Devised with two heuristic cues—the authority of the source and other users' opinions—each of our nudges designates a particular level of credibility of news content on social media. These two heuristics comprise both external authoritative sources of information and social interactions of the crowd. Among the three nudges, both *Reliable* and *Questionable* are applied to information originating from mainstream sources on Twitter, while *Unreliable* makes posts from non-mainstream sources less visible (see Figure 2). Here, *Questionable* and *Reliable* differentiate between mainstream news items that raised questions in their Twitter replies compared to those that did not. Questioned items are highlighted in yellow to warn of the potential controversy in evolving news stories from mainstream media, while those lacking questions are highlighted in green, signifying their reliability. By directing users' attention to the two heuristics, NudgeCred assists users in making meaningful news credibility evaluations.

To demonstrate our nudge-based approach, we built NudgeCred as a Chrome extension [1]. We followed an iterative design process. We first tested our initial design of NudgeCred by conducting a formative study with 16 university students and 36 Amazon Mechanical Turk workers. Results from the formative study helped us refine our design and suggested three confounds—political ideology, political cynicism, and media skepticism—that may restrict impacts on users' credibility perceptions. We then conducted two sets of experiments using our final design: Study 1, a controlled experiment to examine the impact of the nudges with a representative US population (n = 430); and

[1]How NudgeCred works: https://www.dropbox.com/s/2mt4tpdxebccokt/nudgecred_cropped.mp4

Study 2, a qualitative field deployment with Twitter users ($n = 12$) to gain insight into how we can improve NudgeCred's design. Analyzing users' credibility responses from Study 1 revealed that the *Unreliable* nudge significantly reduced users' perceptions of credibility for non-mainstream news sources. For *Questionable*, users in the treatment group rated news tweets with questions as less credible than the users in the control group, and those without questions as more credible. We did not find any effect of users' political ideology, media skepticism, or political cynicism on the effects of nudges. These outcomes suggest that NudgeCred worked irrespective of these confounds. Results from our field deployment (Study 2) show that NudgeCred improved recognition of news content and elicited attention towards all three nudges, particularly *Questionable*. Participants also suggested additional design considerations, such as incorporating heuristics that users would trust, applying nudges to share buttons, and using nudges to distinguish news genres and biases. To conclude, we offer design directions for news credibility nudging by exploring transparency-mode of thinking nudge categories, other heuristics and nudging methods from prior literature. Overall, our contributions include:

- A novel approach using heuristic cues to nudge users towards meaningful credibility assessment of news items in social media.
- A quantitative evaluation of this approach by examining users' credibility perception while considering three confounds—political ideology, political cynicism, and media skepticism.
- A qualitative understanding of the opportunities and challenges of this approach in designing credibility nudges.

2 RELATED WORKS

2.1 Designing for Information Credibility

At present, social media platforms are taking three approaches to combat misinformation—removing misinformation, reducing their reach, and raising awareness [102, 103]. The first line of action falls

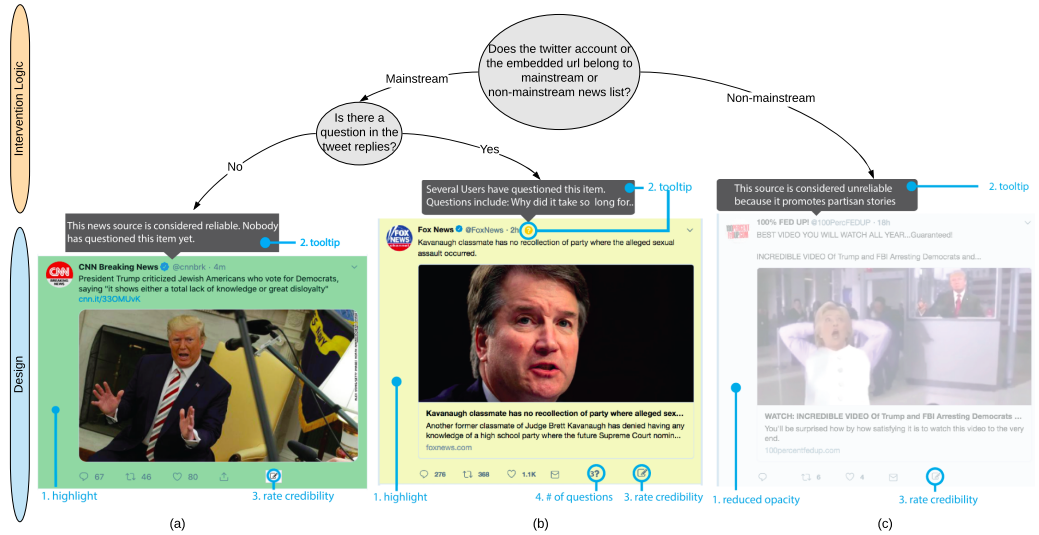


Fig. 2. Our nudge design: [Top] A decision tree shows the intervention logic and [Bottom] three nudge designs. (a). The *Reliable* nudge on a tweet from CNN Breaking News without questions in its comment thread. (b). The *Questionable* nudge is applied to a tweet with questions from Fox News, a mainstream media outlet. (c). The *Unreliable* nudge is activated on a tweet from `100PercentFedUP.com`, an extremely biased, non-mainstream website. The numbers indicate: (1) a change in background, (2) a tooltip message shown when hovered over, (3) a button to open a survey questionnaire for users to rate the credibility of the news tweet, and (4) a button to show more questions in the comments.

under the practices of crowdsourced (in-house and community-driven) and technology-assisted moderation [47, 49, 111] by enforcing established community guidelines [17, 103]. The second approach involves reviews from fact-checking services followed by downranking [18, 68] and the application of warning/correction labels [81, 104, 132]. The third approach largely focuses on contextualizing misleading content through design interventions, such as providing source transparency [34, 113, 124, 127], prioritizing content from trusted authorities [46, 101], and showing related news stories from various sources [67]. Some of these interventions also target particular issues (e.g., voting [105]) or particular interactions (e.g., message forwarding [115]). While contextualizing with additional information has its benefits, producing unsorted information only adds to the confusion [83]. In this regard, we simplify design cues to aid users in distinguishing news coming from mainstream and non-mainstream news sources.

Aside from these platform-led efforts, researchers have also taken up the challenge of designing tools to aid in assessing information credibility. These works span several approaches, including fact-checking systems, interventions, media literacy programs and games [24, 31, 61, 70, 100]. There are multiple scholarly efforts for computationally assessing content credibility [23, 41, 71, 80, 97]. There are some scholarly works on establishing appropriate credibility signals for online content, as well as on designing guides for labeling manipulated media [50, 107, 139]. Some works examine particular crowd-led credibility labeling, including ratings by partisan crowds and the relationship between ratings from crowds and experts [10, 11, 79, 92]. Compared to actively seeking labels from the crowd, our work uses both authoritative sources of information as well as passive crowd interaction with content for labelling content credibility in NudgeCred. This combination helps us overcome the scale issue regarding recruiting quality crowd workers for active labelling.

Scholars have employed multiple types of messages as interventions against misinformation, including theory-centered messages [26], warning messages [19, 89], corrective messages [36, 53, 95], and opposing argument messages [30]. Studies examined the efficacy of interventions in various genres of news, including public health [90, 95] and politics [89]. Some research examined the effectiveness of interventions across countries [40]. Others examined effects for interventions across time by offering real-time correction versus delayed retraction [36, 53]. Real-time correction tools utilize various strategies, including mining databases of well-known fact-checking websites (such as, Snopes and PolitiFact) or crowdsourcing fact-checking. Paynter and colleagues looked into how to strategize corrections by combining several existing techniques (e.g., salience of a graphical element in a warning); they call this approach "optimized debunking" [86]. Some suggest that while corrections can be effective, they can also backfire by inadvertently provoking users into attitude-consistent misperceptions [91]. However, others were unable to replicate such backfiring effects [136]. Warnings about information credibility have been more successful than corrections and are not prone to the same backfiring effect [15]. While there are various methods (e.g., warnings, reminders, and default options [121]) to operationalize nudges, we utilize warnings in NudgeCred.

2.2 Nudges to Steer Human Behavior

The concept of *nudges* has been frequently used to steer civic behavior for achieving important societal goals [42, 120, 122]. The idea stems from behavioral economics and psychology, which define it as a "choice architecture" that encourages citizens to act in a certain way while allowing them to act in other ways, thereby being a favorable alternative to imposing mandates [52]. Such approaches have been highly effective in areas such as environmental protection, financial regulation, and anti-obesity policy [42, 120, 122]. In online settings, technology-mediated nudges have been applied for such purposes as encouraging better password management, improving mobile privacy, and encouraging frugal shopping [7, 8, 54]. Comparatively, nudges regarding online news is getting traction recently in works such as Pennycook and colleagues' "accuracy nudge" (an accuracy

reminder) and Nekmat's "fact-check alert nudge" [82, 90], who investigated impact of nudging on misinformation sharing intention. This work not only extends existing line of research by employing heuristics to assist credibility judgment, but also shows a method of devising such heuristic cue design.

2.3 Heuristic Cues for Credibility

Cognitive psychologists have long argued that when information overload occurs—as it typically does in online social media environments—humans turn to the cognitively effortless route of peripheral processing [25, 93]. While existing HCI works focus extensively on reflective processing, we use automatic or peripheral processing in this study [1]. Peripheral processing means that they depend on heuristic cues, such as the attractiveness of the source or the font color, to evaluate message content [93]. Communication researchers, in response, have offered a well-established list of technology-mediated heuristic cues, highlighted in the MAIN model, which attempts to explain how certain cues influence users' perception of credibility in online contexts [117]. The MAIN model suggests that four technological affordances influence perceptions of credibility: **M**odality, **A**gency, **I**nteractivity, and **N**avigability. These technological affordances trigger various peripheral cues by which users then judge the credibility of online content. For example, the agency affordance focuses on how users perceive source information in computer-mediated contexts. Often, the perceived agent or source of authority can be the machine, the user themselves, or the perceived authors of information on a particular website. For online news, agency is often attributed to the message's source, and these sources can trigger the *authority* heuristic—the perception that the source is an expert on the subject matter [118]. Similarly, information surrounding a message, such as ratings and recommendations, may also provide contextual information. For example, when a group of users likes or shares a news article on social media, the action signals that the group deems the information trustworthy. This signal, in turn, can influence users' perception of the information's credibility while serving as a *bandwagon* heuristic [117, 119]. In summary, the space of all possible heuristics under the four affordances is vast, offering us numerous possibilities for designing credibility nudges. Among these heuristics, we utilize *authority* and *bandwagon* heuristics in our nudge design. We discuss the remaining design possibilities later (see section 8).

2.4 Factors Affecting Credibility Perception: Partisanship, Attitude towards Politics, and Media

Scholars have found numerous factors that may influence information credibility. Based on findings from our formative study (discussed in section 3), we contextualize our research on three behavioral confounds—partisanship, attitude towards politics, and attitude towards media. Historically, scholars have failed to reach a consensus on the role of partisan bias in credibility perception. Earlier work suggested that users perceive unbiased information as more credible compared to one-sided information [4, 87]. Compared to this result, other research found that users would perceive news conforming to their own attitudes as more credible than unbiased news [28, 69, 76]. For this reason, users with strong partisan biases might even resist nudges on attitude-challenging news items, rather than be influenced. Sunstein hypothesized that a considerably large number of people evaluate nudges based on whether they approve of the underlying political objective, naming this "partisan nudge bias" [122, 125]. Hence, we made sure to test our nudge design across a population with balanced partisan affiliations, allowing us to measure the effects of partisanship.

Similar to users' partisan attitude, users' media skepticism can influence their perceptions of the credibility of mainstream and non-mainstream content. Media skepticism is "the feeling that the mainstream media are neither credible nor reliable, that journalists do not live by their professional standards, and that the news media get in the way of society rather than help society" [130].

Scholars have found that media skepticism is negatively associated with exposure to mainstream media and positively associated with non-mainstream media [129]. Media is also generally blamed for its role in enhancing institutional distrust by depicting most governmental policies negatively and causing cynicism towards politics [20]. Studies have also demonstrated that users with high media skepticism and political cynicism rated citizen journalists as more credible than mainstream ones [22]. Drawing from these works, we examine and provide the first empirical evidence of the effects of political ideology, media skepticism, and political cynicism on credibility nudge.

3 FORMATIVE STUDY

We designed NudgeCred in an iterative fashion. Initially, we built a prototype and conducted a pilot study to evaluate it [2].

Method. We built our prototype as a Chrome extension with two types of nudges (described in section 4.2); namely, *Questionable* (tweets highlighted in yellow, which indicate caution) and *Unreliable* (tweets that are less visible). This extension would alter the Twitter homescreen in real-time when users visit them. As mentioned in Figure 2, users could click on a survey button added by the extension. Clicking the survey questionnaire button would open a pop-up overlay comprising our study measurements for credibility. We discus them further in section 5.1.2 (refer Figure 3 to see how it looked). With this setup, we conducted a study with 52 participants from Amazon Mechanical Turk (n=36) and the university (n=36) [12]. For recruitment from the university, we used a university portal available for participant recruitment. For the MTurk users, we used MTurk portal with some filtering conditions, such high rate of work acceptance (>95%), over 18 years of, US resident and familiarity with Twitter. In a pre-study survey, we collected users' demographic details, such as gender and political leaning. Participants were divided into two groups of treatment (seeing tweets with nudges) and control (not seeing any nudge). In a 2-week study period, we asked our participants to rate the credibility of three to five tweets from their Twitter feeds every day. We did so by reminding them everyday to spend around 20 minutes on Twitter by completing an MTurk HIT. Afterwards, we reached out to 16 users—8 control and 8 treatment users—to get feedback on our design where 8 of them finally agreed. In all studies, we compensated our participants adhering to Federal minimum wage requirements ($7.25).

Result. In our study, we hypothesized that users in the treatment group would rate tweets with both *Questionable* and *Unreliable* nudges as less credible compared to users in the control group. A Mann-Whitney U test on the credibility ratings showed that our hypothesis was true for *Unreliable* nudge (avg. cred. (Control) = 0.51, avg. cred. (Treatment) = 0.43 and $Z = 210236, p < 0.001$, Cohen's $d = 1.291$). However, we found the opposite for *Questionable* nudge, i.e., the treatment group rated those tweets as more credible than the control group (avg. cred. (Control) = 0.67, avg. cred. (Treatment) = 0.71 and $Z = 502140, p < 0.001$, Cohen's $d = 0.188$). Furthermore, in our post-hoc analyses, we found that for Republican users the effects of nudges were not significant.

To make sense of the discrepancies in our quantitative result, we conducted interviews followed by a thematic analysis. We identified three themes in the interviews. First, when asked which news organization users follow, participants showed a trend of interest in ideologically aligned news sources. While a majority of Democrats mentioned mainstream sources (e.g., CNN, NBC, the New York Times, and the Washington Post), most Republicans named a mixture of mainstream and non-mainstream sources (e.g., the Wall Street Journal, Fox News, Joe Rogan, and Candace Owens). This trend led us to assume that our intervention may be less effective if it contradicts users' political stances. Second, we found several hints that *cynicism towards politics* and *media*

[2] All of our studies have been approved by our Institutional Review Board.

skepticism can influence the impact of nudges. For example, one participant suggested that he prefers news without biases which mainstream media does not do anymore. Another (Republican) participant expressed frustration that she had to stay away from discussing politics on social media, as she often ran into arguments with others. If Republicans are indeed more skeptical of mainstream media on the whole, and also equally mistrusting of social media platforms, then our intervention could be perceived as yet another attempt by social media to integrate ideologically motivated interventions into their news feeds. Therefore, we decided to examine whether these sentiments of media skepticism and political cynicism adversely affect the interventions. Third, consistent with our quantitative result, we found the opposite of the expected reaction to the *Questionable* intervention. For example, a participant responded: *"I found that these tweets [with Questionable intervention] seem ... more accurate than things that I normally read"*. This conflicting reaction may have stemmed from the lack of a clear hierarchy, i.e., the absence of nudges on more credible news tweets. Subsequently, we revised our design with a third nudge called *Reliable* (tweets highlighted in green to indicate reliability). These findings suggest that our initial prototype did not adequately support better news credibility judgments by users, and informed us to consider three confounds (users' political ideologies and attitude towards politics and media) in evaluating our system.

4 DESIGNING NUDGECRED

4.1 Design Guides

To design nudges with heuristic cues, we employ design guides from two strands of literature: the nudge perspective and the heuristic perspective.

4.1.1 Nudge Perspective. To design effective nudges, the literature suggests two primary aspects to consider: the mode of thinking involved (automatic vs. reflective) and the degree of transparency (transparent vs. non-transparent) [43].

Mode of Thinking: Cognitive psychologists developed *dual process* theories, a set of psychological theories for understanding human decision-making. These theories describe two main modes of cognition: *automatic* and *reflective* [33]. The automatic mode is fast and instinctive. It uses prior knowledge or past repeated behavior and minimal cognitive capacity to decide on actions. Reflective thinking, on the other hand, is slow and effortful. It uses greater cognitive capacity to make a goal-oriented choice by critically examining the effects of choices before selection.

Transparency: Scholars introduced epistemic transparency (i.e., whether users would understand the purpose of a nudge) to divide existing nudge designs into two categories: transparent and non-transparent [43]. Thaler and Sunstein adopted transparency as a guiding principle for nudges [126]. This is because of the concern that a designer may manipulate people into their own preferred direction using systems for behavioral changes.

Using the combination of these two dimensions, Hansen and Jespersen grouped existing nudges into four categories: reflective transparent, automatic transparent, reflective non-transparent, and automatic non-transparent [43]. In designing technology-mediated nudges for credibility, we pick one quadrant from these categories: *transparent* nudges with the *automatic* mode of thinking. We chose the automatic mode as it requires less cognitive effort to process information, especially given the information overload in social media and the instant nature of media consumption. Scholars in the past argued that use of automatic mode over reflective mode for design could address two potential problems–lack of motivation and lack of ability–that typically restrain users from performing tasks such as critically evaluating credibility [1]. Furthermore, our design does not prevent users from critically reflecting on the news content. We chose the *transparent* design to explicitly reveal the motives behind it. We later discuss the potential for nudge designs in the remaining three quadrants (see section 8).

4.1.2 Heuristic Perspective. This work applies heuristics to design nudges for social media in order to enhance users' perceptions of the credibility of news. Cognitive psychologists have proposed models of how effective heuristics work [58]. One of the models, called *Fast and Frugal Heuristics* suggests that users should be able to make inferences using "fast, frugal, and accurate" heuristics when faced with environmental challenges (e.g., information overload) [128]. According to Todd et. al., simple heuristics work when they follow two principles: they exploit the structure of the environment and are robust. In social media, structures include sources of news items, popularity (indicated by the number of shares or replies), and the way that information is organized by time and personal interactions. They argued that heuristics that exploit existing structured information can be "accurate without being complex" [128]. Another success criteria for heuristic design is the robustness of the decision model. A computational strategy utilizing a limited set of information can yield more robustness [128]. Employing these principles, our design includes only two heuristics, outlined below. These heuristics seem useful to users to investigate misinformation [37].

- *Authority Heuristic*: We limit the source of news to a handful of known organizations followed by a binary classification of the organizations.
- *Bandwagon Heuristic*: We utilize the conversational structure (or replies) of the environment as an indicator of credibility assuming a skew in the reply distribution.

4.2 Outlining the Design

Our design of NudgeCred is built on the idea of applying subtle heuristic cues in certain contexts in social media. It is powered by three types of socio-technical interventions—*Unreliable*, *Questionable*, and *Reliable*. Using the principles of fast and frugal heuristic design, our design uses a two-level decision tree with two heuristics (see figure 2). They are triggered based on whether a news tweet originates from an official authority. Thus, the first step of our tool design relies on the *authority heuristic*. Communication scholars have long argued that revealing the official authority of content results in applying the authority heuristic in credibility judgments [117]. We apply the authority heuristic by differentiating between mainstream and non-mainstream news tweets. We do not apply nudges to tweets that do not come from mainstream and non-mainstream sources. We opt to use source-based credibility annotation due to the challenging nature of annotating article-level credibility. While we may flag some accurate articles from non-mainstream media in this method, other work demonstrated that this number could be few (14%) compared to the accuracy (82%) [110]. For such false-positives, users still have the opportunity to fact-check them.

To flag inaccurate content from mainstream media, we apply another criteria of whether someone replied to a mainstream tweet with a question. Literature suggests that such questions, depending on users' prior knowledge of the subject matter, can instigate curiosity and motivate them to investigate to a varying degree [65]. We employ this property by showing the number of questions as a *bandwagon heuristic* ("if others question this story, then I should doubt it, too") on mainstream news tweets. Thus, our study had three nudges: mainstream news tweets that did not have questions raised about their content (*Reliable*), mainstream news tweets that had questions raised about their content (*Questionable*), and non-mainstream news tweets (*Unreliable*). To employ epistemic transparency in the design, our design includes a tooltip with the reasoning behind each nudge. Figure 2 shows the overall design. Before delving into each intervention, we propose a classification of news sources that enables our application of the *authority heuristic*.

4.2.1 Classifying Authority of News. Our nudge-based interventions work on two types of news sources: mainstream and non-mainstream. In journalism and communication literature, the term "mainstream media" lacks an official definition. For the purposes of our study, **mainstream news sources** are sources recognized as more reliable in prior scholarly work. Opting for a heuristic

Mainstream Source		
The Economist	CNN	The Blaze
New York Times	NPR	BBC
Washington Post	MSNBC	Fox News
Chicago Tribune	WSJ	Politico
New York Post	Newsday	NY Daily

Table 1. Example sources in our *mainstream news* category.

Website	Category	Inaccuracy Type Message
abcnews.com.co	Fake news	misinformation
breitbart.com	Extreme Bias	partisan stories
americantoday.news	Rumor Mills	rumor
infowars.com	Conspiracy Theory	conspiracy
rt.com	State News	state propaganda

Table 2. Example *non-mainstream news sources* and their categories of reporting inaccuracy. The tooltip messages read: "This source is considered unreliable because it promotes <InaccuracyType>".

approach, we use such existing literature to create a reliability measure which may later be replaced; this is not our primary contribution. In this approach, the first two authors iteratively collected a list of mainstream news websites by referring to two prior works, including Pew Survey and NPR [78, 109]. Next, we refined our list with the help of our in-house journalism and communication media expert by referring to the most circulated and the most trusted news sources [27, 78], subsequently removing a news aggregator (Google News) and a local source (AMNewYork). Table 1 shows a sample. Every source on our final list of 25 news sources follows standard journalistic practices in news reporting [5].

For **non-mainstream news sources**, we refer to `Opensources.co`, a professionally curated list of websites known to spread questionable content [84]. Each source in this list is categorized based on its level of information reliability (e.g., 'extreme bias,' 'rumor,' 'clickbait'). The curators manually analyzed each source's domain-level characteristics, reporting and writing styles before assigning a particular category. From this list, we remove the 'politics' and 'reliable' categories to retain sources which were explicitly labeled as promoting unreliable news, a total of 397 sources spanning 10 categories. Table 2 shows a sample from this list. We do not intervene in the rest of the sources that do not fall into these two categories. Using this notion of mainstream and non-mainstream news sources, we apply three nudges.

4.2.2 Three Nudges. The ***Unreliable*** nudge detects whether a tweet comes from an unreliable authority. Our design applies this nudge by examining whether a tweet from a user's feed originates from a non-mainstream news site and subsequently reduces the item's opacity, rendering it harder to read. We call these tweets unreliable non-mainstream tweets (T_U) [See (c) in figure 2]. To instigate epistemic transparency, *Unreliable* provides an explanation of its action through a tooltip message: "This source is considered unreliable because it promotes <InaccuracyType>." Table 2 shows the list of <InaccuracyType> messages based on the source and its category in *opensources.co*.

The ***Questionable*** nudge is applied to mainstream news tweets (T_Q) when at least one question is raised about the information in the corresponding Twitter reply thread[3]. Prior studies suggest that less credible reports on an event are marked by questions and inquiries [80, 140]. To detect questions, our algorithm is kept intentionally simple. Our algorithm looks for "?" mark to identify questions—a simple but transparent method that is understandable by users. It is worth noting that our focus is not to develop the most sophisticated algorithm to detect questionable news, rather testing the effectiveness of nudge in credibility assessment. In that regard, using the simple heuristic serves its role and have benefits in simplicity and transparency for users to understand. While investigating advanced natural language parsing methods to identify relevance of the questions to the news article or more advanced machine learning techniques to detect questions is worth looking into [62, 138], such investigations would require significant work, perhaps amounting to a separate full contribution. Hence we leave those as future paths to pursue. Instead our approach works as a minimum baseline to identify questions. To make users aware of these questioned mainstream

[3]Though Twitter has recently rolled out a threaded reply structure, at the time of the study, it did not exist. Thus, we only took direct replies into account.

Source Type	Twitter Account	Political Bias	Source Type	Twitter Account	Political Bias
Mainstream	CNN Breaking News	Left	Non-mainstream	Daily Kos	Left
Mainstream	NY Post	Right	Non-mainstream	Breitbart	Right
Mainstream	Politico	Center	Non-mainstream	Zero Hedge	Conspiracy

Table 3. Example news sources and their political biases.

tweets, the *Questionable* nudge is applied by changing the background color of the tweet to yellow while showing the number of questions (see (b) in figure 2). By showing this number, we promote a collective endorsement that multiple users have doubts about this news [119]. Additionally, a tooltip message offers transparency by explaining the reason behind the nudge activation. For T_Q, the tooltip message follows the format: "Several users have questioned this item. Questions include: <first reply tweet with a question>" (e.g., for a tweet containing a news report with missing details, such as time of an event, a reader may have replied to ask: "When did this happen?"), thus directing further attention to other users' comments in an effort to stimulate the *bandwagon heuristic*. The bandwagon effect has been demonstrated to be powerful in influencing credibility judgments [57].

The ***Reliable*** nudge is triggered when the source of the news tweet is an official, mainstream source and was not questioned in the replies; specifically, reliable mainstream tweets (T_R) were emphasized with a green background highlight (see figure 2 (a)). A tooltip message is formatted for T_R as follows: "This tweet seems more reliable. Nobody has questioned this item yet." The colored highlights and the corresponding tooltip messages create contrast within the mainstream news tweets, helping users navigate them better. Note that we included this nudge in response to the findings from the formative study.

5 STUDY 1: EVALUATING IMPACT ON PERCEPTIONS OF CREDIBILITY IN A CONTROLLED SETTING

To evaluate our design, we conducted two studies. Study 1 evaluates impact on credibility perception in a controlled setting while Study 2 is a field deployment. In Study 1, we examine three research questions on the effect of nudges on users' credibility perceptions in a controlled setting simulating a Twitter feed with multiple tweets for each nudge.

RQ1. *Can heuristic-based design nudges on an online social news feed help users distinguish between reliable, questionable, and unreliable information?*

RQ2. *Do users' partisan attitudes affect their responses to credibility nudges?*

RQ3. *Do users' political cynicism and media skepticism affect their responses to credibility nudges?*

5.1 Method (Study 1)

5.1.1 Selecting news tweets for a controlled environment. For this study, we simulated a Twitter feed with a fixed set of tweets which would be shown to every user. To simulate a realistic Twitter feed, we selected news sources from our previously compiled list of mainstream and non-mainstream sources, and then selected several tweets from each source (see Table 1 and 2). We used a balanced approach to select sources with left-wing, centrist, and right-wing biases. For bias categorization, we used `mediabiasfactcheck.com`, a source used in other scholarly works [55, 72]. For each news source under each bias category, we first found the source's Twitter handle. We retained the Twitter accounts that had the greatest numbers of followers[4] (a mark of popularity of the news source on Twitter). Table 3 shows sample Twitter accounts with their perceived political biases. For each source, we selected the tweet within the last 48 hours that had the highest number of shares. With three nudges working across three political leaning categories, our feed comprised 9 tweets (3 political leanings × 3 nudges). Appendix A shows several tweets from this list. To add variation, we created another set of 9 tweets using the second most followed Twitter accounts from our list

[4]We excluded Fox News' Twitter handle due to its inactivity for several months until the time of the study.

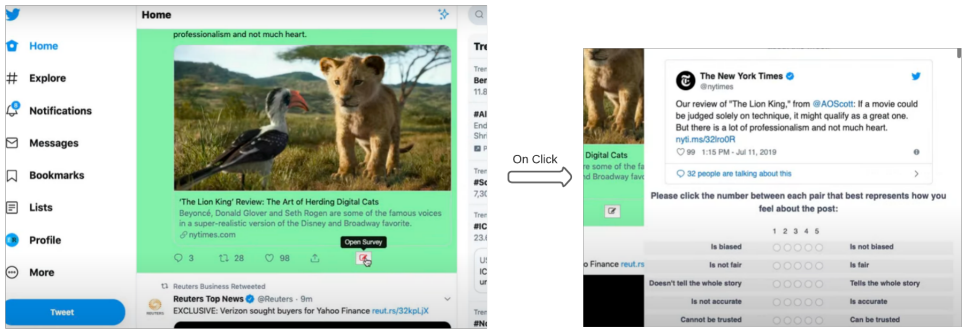


Fig. 3. Screenshot of how clicking on the survey button would pop open the five-item credibility questionnaire.

Item	IRR
Is/not biased	0.83
Is/not fair	0.79
Does/not biased tell the whole story	0.80
Is/not accurate	0.79
Can/not trusted	0.79

Table 4. IRR of the five-item questionnaire on credibility in the formative study.

Pol. Cyn.	1. Elected officials put their own interests ahead of public's interest
	2. It seems like politicians only care about special interests
Med. Skep.	1. The media provide accurate information
	2. The media provide trustworthy information
	3. The media deal fairly with all sides
	4. The information provided by the media needs to be confirmed

Table 5. Items used in measuring political cynicism and media skepticism. We used a five-point Likert scale (Strongly Agree – Strongly Disagree) with a "Don't know" option.

of news sources, resulting in a second feed for our controlled environment. Users were randomly shown one of the two feeds, totaling 9 tweets in each case. To evaluate our RQs, we needed to measure users' credibility perception, political ideology, political cynicism, and media skepticism.

5.1.2 Measuring News Credibility, Political Ideology, Political Cynicism & Media Skepticism. We used a five-item questionnaire by Meyer et. al. [77] to measure users' perceptions of credibility for every news tweet (see Figure 3). In our formative study, we found this measure had a high Cronbach α ($\alpha = 0.95$) and individual inter-item correlations (see Table 4), showing a high level of internal consistency. To capture partisan attitudes, we survey participants for their political ideology on a seven-point Likert scale ranging from "strong Republican" to "strong Democrat". We survey participants on media skepticism and political cynicism using a validated questionnaire from journalism scholarship (see Table 5) [22]. For both variables, we average the responses across questions and use the median to create a binary response variable with values "low" and "high." Note that we had to revert the first three media skepticism questions before averaging.

5.1.3 Recruitment. Our study participants were recruited starting the third week of July 2019 and spanning a period of three weeks. We required three qualifications for user participation: (1) age of 18 or older, (2) US resident, and (3) familiarity with Twitter. This choice of US population was purposeful due to the difficulty in measuring our confounds across global population. Users'

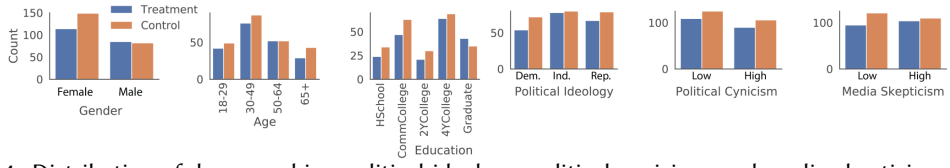


Fig. 4. Distribution of demographics, political ideology, political cynicism, and media skepticism in our participants in Study 1.

political leaning has different meanings in different countries (e.g., political left-right are different in the US and Europe). Similarly, levels of skepticism/cynicism might vary by country. We focused on US-population due to the availability of well-established measurements for our confounds from the communication literature [22, 77]. We recruited 430 users from *Qualtrics*, well-balanced by partisan affiliations. Figure 4 shows their demographics. This sample is mostly balanced across gender, age, and education, with a slight skew toward females.

5.1.4 Study Procedure. We presented our participants a set of tweets collected right before the start of recruitment. We chose this approach because studies have shown that there is a lag in terms of the amount of time media coverage takes to influence public opinion, with some exceptions (e.g. mass shootings) [99]. As a result, we anticipated that participants would be least likely to be familiar with the most current tweets. Participants were randomly assigned to either the treatment or the control group, with a quota check to ensure balanced allocation across political ideology. To counter order effects, we presented tweets in random order. We added attention checks–questions with options reversed–right before the Twitter feed. Taking recommendations from Qualtrics, we also discarded participants who spent less than six minute to respond–the half of the median time spent by users in a soft-launch of 50 users. Participants saw each tweet and answered the questions for that item before scrolling down to the next one. This approach reflects a natural setting of modern social media sites, where users browse feeds in real-time, click links in-situ, and the same post usually do not appear again at a later time point. To reduce response bias, we framed the questions to ask for credibility of the items (e.g., how do you feel about the post?) instead of the effects of nudges (e.g., how does the nudge affect your perception?). The unexpected effect on *Questionable* in our formative study suggests a lack of response bias.

5.1.5 Method of Analysis. We initially perform mean comparison with Mann-Whitney U-test. However, note that each user saw multiple tweets with each intervention. To model such repeated measurements for the same intervention, we further use a mixed-effects logistic regression.

$$\mathbf{y} = X\beta + Z\mathbf{u} + \epsilon \tag{1}$$

In Eq. 1, the response variable (credibility score) ($\mathbf{y}$) is the dependent measure for our experiment. While fixed effects (X) are the independent measure, random effects (Z) are the variables repeated in multiple observations; that is, tweets. The residual (ϵ) is the error in fitting. Finally, β and $\mathbf{u}$ are the coefficients of fixed and random effects, respectively. We used an **R** implementation of a linear mixed-effects regression model, *lme4.lmer*, on our dataset [14].

Dependent Variable: Our dependent measure is the credibility score, a continuous variable computed by averaging the five credibility question responses (see Figure 4) followed by standardization. We perform robustness checks by rerunning mixed-effects ordinal logistic regressions on each of the five credibility questions. We find no significant differences in the resulting model coefficients, suggesting sufficiency in modeling the credibility score as continuous.

Independent Variables: The independent variables related to RQ1 include main effects and interaction effects derived from the two experimental conditions: users' group (control or treatment) and intervention type (T_R, T_Q, T_U). For RQ2 and RQ3, we examine three variables, including political ideology, political cynicism, and media skepticism. For the sake of analysis, we map political ideology, measured on a seven-point Likert scale, to three groups consisting of Democrats, Independents, and Republicans. Following a prior scholarly work [22], we used the median score across all the questions in each variable (political cynicism and media skepticism) to split the participants into two groups. In our representative US sample, the median for media skepticism was 2.75 ($\alpha = 0.72$, $M = 2.48$, $SD = 0.98$) and the median for political cynicism was 4.00 ($r = 0.53$, $M = 4.09$, $SD = 0.81$). Similar to Carr et al., we considered values greater than the median as high on that category and

	Control Avg. Cred.	Treatment Avg. Cred.	Bet. Subj. MWU-test (Cohen's d)
T_R	0.62	0.67	187488.0(0.162)***
T_Q	0.58	0.55	198180.5(0.072)*
T_U	0.46	0.37	171763.0(0.296)***
n	693	597	

Table 6. Mann-Whitney U test results for Study 1. Here, 'n' denotes the number of tweets rated in each condition. Avg. Cred. is the mean of 'n' credibility scores; * $p < 0.05$, *** $p < 0.001$.

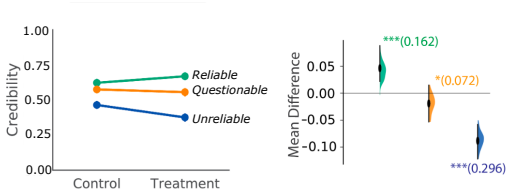


Fig. 5. Shows interaction effects between user groups and nudge types in Study 1. The numbers inside the brackets are the effect sizes, Cohen's d.

vice versa [22]. In other words, media skepticism of 3.00 would be labeled high media skepticism, while 2.75 would be labeled low media skepticism. We also include the political leanings of the news sources used in our tweet selection procedure as an additional independent variable.

Control Variables: Prior studies indicate that the level of interest in a news story can influence users' credibility assessment [74]. Therefore, we include participants' interest in a tweet as a control variable, measured on a five-point Likert scale ranging from low to high interest. Other control variables include users' demographics, such as gender, age, education, and Twitter usage frequency.

5.2 Results (Study 1)

5.2.1 RQ1: Effect of the Nudges. For RQ1, initially we investigated our data using mean comparison. Table 6 shows the mean values and Mann-Whitney U test results of our experiment and Figure 5 shows corresponding interaction plots. Users in the treatment group rated the credibility of non-mainstream news tweets (T_U) significantly lower than did users in the control group, suggesting the effectiveness of our intervention (Z=171763, p<0.001, Cohen's d=0.296). Additionally, treatment users rated mainstream news tweets without questions (T_R) as more credible than corresponding control users (Z=187488, p<0.001, Cohen's d=0.162). Our participants showed significant decrease in their rating of mainstream news tweets with questions (T_Q) (Z=198180, p<0.05, Cohen's d=0.072).

Our experimental setting with each user rating multiple tweets prompted us to further analyze the data using a series of mixed-effects regression models. Table 7 shows this analysis. To determine the effects of the experimental conditions, our base model includes news source type, group assignment, and their corresponding interactions. We find that tweet type—mainstream or non-mainstream—is strongly correlated with a tweet's credibility score. Non-mainstream tweets are generally rated less credible than mainstream news sources with a small effect size (β=−0.16, p<0.01 and Cohen's f=.07). This result suggests that users can differentiate between mainstream and non-mainstream tweets even without our nudges. However, treatment users (those who received nudges) generally rated tweets as slightly more credible than control users (β=0.04, p<0.01, Cohen's f=0.04). Nevertheless, there is an interaction effect between tweet type and user group. Treatment users scored non-mainstream tweets lower than control users with a medium effect size (β= − 0.10, p<0.001 and Cohen's f=0.16). Treatment users also rated mainstream tweets with questions as less credible than did control users (β= − 0.06, p<0.001, Cohen's f=0.16). The decrease in the credibility perception scores of both mainstream questioned tweets (T_Q) and non-mainstream ones (T_U) suggests that our nudges can help users consume these news items as less credible, thereby answering RQ1.

5.2.2 RQ2: Effect of Political Ideology. To answer RQ2, we examine the political ideology variable in our politics and media regression model. Politically Independent users serve as the point of reference for this variable. We find that Democrats generally rated all tweets slightly higher in credibility than did Independent users (β=0.03, p<0.05, Cohen's f=0.06), whereas Republicans rated them slightly lower than Independents (β= − 0.03, p<0.05, Cohen's f=0.06). Due to this

	Base Model		Politics & Media Model		3-Way Interaction Model	
	β	SE	β	SE	β	SE
(Intercept)	0.29***		0.27***		0.32***	
Control Variables						
Gender (Male)	0.03*	0.05	0.03*	0.05	0.04	0.07
Education	-0.00	-0.01	-0.01	-0.03	-0.01	-0.04
Age	0.00	0.01	0.01	0.03	0.01	0.04
Social Media Usage	0.01	0.01	0.01	0.02	0.01	0.02
Interest in the Tweet	0.09***	0.45	0.09***	0.45	0.08***	0.36
Experimental condition						
Type(Mainstream-Question)	-0.03	-0.05	-0.03	-0.05	-0.07	-0.11
Type(Non-mainstream)	-0.16**	-0.26	-0.16**	-0.26	-0.23**	-0.38
Group(Treatment)	0.04**	0.07	0.04**	0.07	0.08	0.13
Type(Mainstream-Question):Group(Treatment)	-0.06***	-0.08	-0.06***	-0.08	-0.11*	-0.14
Type(Non-mainstream):group(Treatment)	-0.10***	-0.13	-0.10***	-0.13	-0.21***	-0.26
Politics and Media						
Ideology(Democrat)			0.03*	0.05	0.01	0.01
Ideology(Republican)			-0.03*	-0.04	-0.03	-0.05
Political Cynicism(Low)			0.01	0.02	0.01	0.02
Media Skepticism(Low)			0.05***	0.08	0.06**	0.11
Account Leaning(Conspiracy)			0.07	0.08	0.06	0.07
Account Leaning(Left)			0.09	0.14	0.09	0.14
Account Leaning(Right)			0.00	0.01	0.00	0.00
Experimental Condition x Other Variables						
Group(Treatment):Gender(Male)					-0.08*	-0.11
Type(Non-mainstream):Interest in the Tweet					0.02*	0.11
Type(Non-mainstream):Group(Treatment):Gender(Male)					0.09*	0.08
Type(Non-mainstream):Group(Treatment):Interest In a Tweet					0.03*	0.10
Adj. R^2 (Marg./Cond.)	.310/.483		.347/.494		.358/.508	
N = 3870					* p<.05, ** p<.01, *** p<.001	

Table 7. Regression models on the credibility score. The base model contains nudge type, user group, control variables and the interaction between user group and nudge type. The politics and media model adds users' political ideology, media skepticism and political cynicism variables to the base model. The 3-way interaction model further includes the interactions of nudge type, user group and other variables with significant main effects in the politics and media model (Gender, Interest in the Tweet, Ideology and Media Skepticism).

main effect, in our 3-way interaction model, we further explore whether political ideology had any interactions with the three nudges and the users' group assignments. However, we find no significant interaction. Therefore, nudges changed users' credibility perceptions irrespective of their political leanings. We discuss these findings later.

5.2.3 RQ3: Effect of Political Cynicism and Media Skepticism. To answer this RQ, we examine two variables (political cynicism, and media skepticism) in our politics and media regression model. Between media skepticism and political cynicism, only media skepticism had a significant effect, where users with lower media skepticism rated tweets as more credible (β=0.05, p<0.001, Cohen's f=0.10). In our 3-way interaction model, we further explore whether media skepticism had any interactions with our key variables of interest—treatment and control groups, and the three nudges. We do not find any significant interaction effects. Therefore, nudges changed users' credibility perceptions irrespective of their attitudes towards politics and media. We elaborate on these findings in the Discussion section.

5.2.4 Effect of Control Variables. We examine whether user demographics and users' interest in a news story had any effect on how they rated the credibility of the news tweet. Across all three models, the effects exerted by our control variables are consistent. Independent of whether a user was assigned to the control or treatment group and independent of the type of news (whether mainstream or non-mainstream) they saw, users provided higher credibility scores when they were interested in a story, with a large effect (base model effects:β=0.09, p<0.001, Cohen's f=0.5). Among demographic variables, male users rated tweets as more credible with a small effect size

(base model effects: β=0.03, p<0.05, Cohen's f=0.03). The remaining demographic variables did not show any significant effect [5].

6 STUDY 2: FIELD DEPLOYMENT

To gain insights into how we can understand and improve the current design nudges for credibility, we conducted a qualitative study. For this purpose, we recruited participants to use NudgeCred on Twitter for five days. This process allowed us to evaluate it in a more ecologically valid environment than Study 1 [16]. Below, we describe the process.

6.1 Method (Study 2)

6.1.1 Recruitment. To recruit users for this study, we used Twitter advertising, following a prior strategy [50]. We devised several targeting mechanisms to promote the advertisement to our desired group, including age range (>=18), language (English), location (USA) and whether users followed top mainstream and non-mainstream news accounts in our list. Initially, we were not successful in getting responses from broader regions, so we iteratively revised the location to target nearby states for effective recruitment. Additionally, we promoted the advertisement within our academic network on Twitter. From 50 interested participants, we recruited 12 participants for the study by filtering our spams and users with less than 100 followers . Overall, our participant group consisted of 5 females and 7 males with an average age of 29.5 (std. dev. = 6.5) and a political tilt towards Democrats (Democrat = 6, Independent = 4, Republican = 2).

6.1.2 Procedure. Followed by an informed consent process, we instructed users to install the NudgeCred browser extension. To promote a natural setting, we encouraged users to use Twitter during the five-day study as they normally would. However, it is possible that Twitter's news feed algorithm may not have surfaced news items on their feed each time. Hence, we also encouraged users to visit the Twitter profiles of some news sources each day to ensure that users experience how NudgeCred works. After five days, we asked them to fill out a post-study survey on demographic details followed by an interview in a semi-structured manner (see Appendix B for the interview questions). To facilitate their responses, we asked users to walk us through their news feed during the interview. Each participant received a $30 gift card for participating.

6.2 Results (Study 2)

We analyzed the interview data using a grounded theory approach [114]. The first author transcribed the audio from the interviews and analyzed the data to come up with a set of themes. These themes were discussed and refined with the other authors. Below, we present our final set of themes.

6.2.1 NudgeCred facilitates more conscious news consumption. Most of our participants (9/12) provided positive feedback on their overall experience with NudgeCred, referring to the design and application. Some participants (U1,U6,U9) particularly mentioned that they liked the bandwagon heuristic with the questions in replies.

> *"It [NudgeCred] quickly highlights. So you know what to look for. Especially when it's a question mark, I do actually open up the comment section and read what the question is."* ***(U9)***

Others liked it because it served as an educational tool to "train the user to thoughtfully think about" news (U4) or because it did not add overhead to their current Twitter experience (U2). Overall, users reported two phenomena. We describe them below.

[5] Additionally, we examined whether there was any learning effect compounded from seeing multiple nudges. To do so, we added the order (from 1 to 9) of the tweets in which participants evaluated their credibility and its interaction with nudge type and user group in our regression models. We found no significant effect of the order.

Improved Recognition of News Content and News Genres: One of the impacts that participants (5/12) mentioned was the perceived difference in the amount of news content in their feed compared to their prior experience. For example, they perceived that there was more news content in their feed than before. NudgeCred even helped some participants pay more attention to the types of content that news sources produced.

> *"It [NudgeCred] really just told me that NPR produces more articles that are opinionated [the participant was referring to an Op-Ed] on Twitter than I thought."* ***(U1)***

The article labeled as *Questionable* made them realize that it was an op-ed article, rather than news.
Attention towards Bandwagon Heuristic: Out of the three nudges, participants (7/12) noticed the *Questionable* tweets highlighted in yellow the most, and the number of questions below them.

> *"I noticed the one [question icon] at the bottom more ... most people who use Twitter a lot ... 9 out of 10 are more likely in tune with the replies/retweets. Usually I use those as a sign of popularity."* ***(U1)***

> *"If I see it [a tweet] as yellow ... I do get the information that this is either ... a lot of people don't like the news article or this article might have controversial or incorrect facts."* ***(U10)***

While users, as U1 indicated, may see traditional retweet or reply numbers as an indicator of popularity, one participant (U10) correctly pointed out the bandwagon cue as an indicator of controversy. Thus, nudges imitating existing designs on social media can be useful.

Overall, these phenomena support that our nudges can improve users' news perception in two ways: (i) with an overall impression of total news items on users' feeds broken down based on the reliability of sources, facilitating better perception on its genres; and (ii) with individual attention towards particular news items.

6.2.2 Concerns in Using Heuristics for Nudging News Credibility. Interviews also revealed two concerns regarding our nudge design. We discuss these concerns below.
Trust Issues with Heuristics: A majority of our participants (7/12) questioned the use of bandwagon heuristic to differentiate *Reliable* and *Questionable* news items. Because audience composition can vary by the source and the topic of a news item, and influence bandwagon heuristic, they were concerned about its disparate impact. One participant pointed out that followers of the New York Times (NYT) are comparatively more diverse than followers of Fox News. Consequently, audiences with an opposing stance on a news report from the NYT may question it. In contrast, Fox News, having a homogeneous audience which mostly supports its reporting, may hardly question their reports. Therefore, our bandwagon heuristic would be skewed based on the audience of a report.

> *"NYT [The New York times] and MSM [mainstream media] in general have a lot more reactions from skeptical readers given the current administration. And to some, the color-coding and the number of questions may indicate that the news is subjective or "fake" when you compare it with other outlets such as Fox News that have fewer reactions on Twitter and a more homogeneous audience."* ***(U7)***

These responses suggest that even though a user may understand the bandwagon heuristic, the heuristic itself may have some shortcomings, which makes it challenging for the user to trust it as a metric for gauging credibility.
Adverse Effects of Nudges: Our participants (2/12) suggested two adverse effects of the nudges. One participant (U11) proposed that users may use the bandwagon heuristic based on like-minded questions as a justification for their attitude-consistent belief in politically opposing news, thus promoting a **confirmation bias**.

> *"If I agree with you and you are questioning an article that I questioned as well ... since you personally agree with me, it confirms my bias against that piece of information."* ***(U11)***

A similar effect has been suggested by scholars in the past, who show that rating mechanisms on online platforms (e.g., Facebook "likes") may guide users' news selection process [73]. If such effects exist, users may become more polarized. Compared to confirmation bias stemming from the existence of a nudge, the **absence of nudges**, mentioned by U4, could also have a harmful effect.

> *"I think the ones that I subconsciously ignore are the ones that have no color at all. If there aren't any flags ... no color blocks, I am more inclined to assume that the content is valid."* ***(U4)***

This participant suggested that the absence of nudges creates an illusion of validity of content without nudges. Indeed, recent research points out the same phenomenon when false reports are not tagged, resulting in a false sense of being validated [88]. One way to address this concern is, again, to be transparent about not being nudged with an additional tool tip message for the news items that are not nudged.

Overall, our participants' concerns suggest that designers need to evaluate two aspects of nudges: (i) How trustworthy the design components of the nudges are (ii) Whether the presence and absence of nudges adversely affect users.

6.2.3 Opportunities for Credibility Nudge Design. In addition to differentiating news credibility, we asked participants what other functions they would like in NudgeCred. Participants suggested improvements in three directions.

Extending the News Source List: Our participants were concerned on the limited set of news sources we considered. They (5/12) suggested that they often see misinformation from non-news entities, including their acquaintances. To allow more news identification, some (2/12) asked us to include local news sources. With our participants following diverse sources for information, our limited news source list was naturally inadequate for them.

Indicating News Genres and Reporting Biases: Suggestions from the participants included distinguishing opinion items from news (3/12) and indicating bias in a report as a nudge (2/12).

> *"Give a notification that say what I am seeing is an op-ed rather than straight facts."* ***(U4)***

> *"Is it possible to state that this news article is biased towards this particular claim?"* ***(U10)***

A recent survey shows that about 50% US adults "are not sure what an op-ed is" and that about 42% of respondents perceive that opinion and commentary are often posed as news in most news articles [96]. Therefore, a significant share of the population may appreciate having nudges that differentiate op-eds from news as well as other indicators of bias stems. Incorporating such attributes in nudging might help users better determine the credibility of news content.

Curbing Misinformation Sharing: To prevent the sharing of misinformation, some participants (2/12) proposed implementing nudges on share (or retweet) buttons.

> *"[when someone clicks the share button] If there is a notification that says this source is not credible then people would be less likely to share it."* ***(U2)***

Research indicates that about 59% of links shared on Twitter have never been clicked [35], i.e., users often share news items without reading them. If nudges can help user determine the unreliability of news from misinformative sources, they might be less likely to share them.

In summary, our participants proposed improvements to our nudge design in three key areas: (i) improving existing classifications by extending the source list, (ii) expanding news classifications of nudges in alternate areas, and (iii) targeting users' interactions with news on social media.

7 DISCUSSION

Below, we elaborate on our results, starting with each research question from Study 1, followed by opportunities and challenges suggested by the participants in Study 2.

7.1 RQ1: Effect of Nudges on Credibility

Our regression analyses in Study 1 revealed that users' credibility ratings were considerably different between the treatment (nudged) group and the control group. While other nudge designs have proven effective in reducing sharing intentions of misinformative content, their effectiveness was shown on a particular news genre, such as COVID-19 and HIV [82, 90]. In contrast, we examined the effects of our nudges on a wide variety of popular news items surfacing over multiple days, thus offering a more generalized result. Our intervention provides a less authoritative approach that gives users simple but transparent information for them to make their own judgments. News feeds, as they typically present limited information on social media, have few features to distinguish content quality. Tweets from bots, friends, mainstream news, and non-mainstream sources are all given equal weight in terms of visual representation in any feed, making it difficult for users to sift through them. Though people are capable of identifying misinformation, social media makes it challenging to make informed analytical judgments [89]. Our results suggest that users might appreciate it if social media sites provide tangible signals to work through this clutter, which is further exemplified by participants' suggestion to differentiate news and op-eds in Study 2.

Apart from facilitating better perceptions of the credibility of news, NudgeCred may also act as a "translucent system" [32]. The theory of social translucence posits that we should aim for systems that make online social behavior *visible* to facilitate *awareness* and *accountability*. Note that our participants in Study 2 suggested improved recognition of particular types of news content on their feed and were more aware of what they were seeing on their feeds. Our nudges on news content that are liked or shared by users' followers or friends could also have similar impacts, wherein users become more *aware* of their peers' news consumption behaviors. When their peers like/share misleading news, one may hold the peers *accountable* by pointing out the misleading content. Besides, after seeing the nudges on unreliable content, users may restrain themselves from sharing such content and reflect on their sharing habits.

7.2 RQ2: Influence of Political Partisanship on Nudge Effects

Our regression results suggest that NudgeCred changed users' perception of credibility irrespective of their political views; that is, there were no interaction effects between political characteristics and the effects of interventions. This result is consistent with recent studies showing the success of interventions in limiting sharing intentions of misinformation, irrespective of users' political affiliation [100, 136]. Although some prior literature argue that citizens may view nudges as partisan policies and may not support nudges when they conflict with users' partisan preference [122], other scholars suggest that this behavior can be countered by reducing partisan cues in nudges [125]. We incorporated this suggestion in our design by showing nudges on news content from all political leanings and nudging news content in both directions (Reliable, Questionable, and Unreliable). However, in practice, users tend to typically follow news based on their partisan preferences [59]. In such a setting, users who follow only alternative fringe sources may see mostly *Unreliable* nudges triggered on their reports and perceive NudgeCred as partisan. One potential design solution is to show the similar news item from reliable sources, with the same partisan view, to help them understand the alternatives.

7.3 RQ3: Influence of Political Cynicism and Media Skepticism on Nudge Effects

In our study, we did not find any impact of political cynicism or media skepticism on nudge effects. This convincing nature of our nudges—that nudges worked irrespective of users' prior media skepticism and political cynicism—is promising. Our result for media skepticism is aligned with a recent work where media skepticism did not affect nudge effects [82]. Research suggests that

media skeptics, despite significant exposure to alternate sources, still seem to have moderate to high consumption of mainstream news [131]. Therefore, our nudges could improve news credibility assessment of both mainstream and non-mainstream sources by skeptics in the wild. Scholars suggest that exposure to fake news mediated by belief in its realism increases political cynicism [9]. Thus, if nudges can reduce belief in fake news, it could help mitigate increasing cynicism towards politics. Furthermore, our nudges can be utilized as an alternative to censorship by social media, thus helping mitigate the concern that social media apply censorship in a disparate manner across different political affiliation, as raised by participants in our formative study.

7.4 Opportunities in Designing News Credibility Nudges

Our field deployment of NudgeCred showed several opportunities in designing credibility nudge in the future. First, participants' attention to the bandwagon heuristic reveals how designers can utilize existing Twitter infrastructure in their design. Though the impact of the bandwagon effect in collaborative filtering has been discussed in the literature [119], it has been underutilized in a news credibility context. Our study suggests that applications similar to ours can act as valuable markers of information credibility. Second, participants seeking nudges on a wider set of sources (e.g., news and non-news sources), and alternate types of taxonomies (e.g., news and op-eds) suggests their need for tools to differentiate information reliability. Comparatively, nudges on tweet actions (e.g., retweet and share) may play a stronger role in curbing the spread of misinformation, as research indicates that sharing without thinking is often weaponized for this purpose [135]. For example, when users click on the share button, activating a "timer nudge"—a visual countdown to complete the action—could make users rethink before they share [134].

7.5 Challenges in Designing Nudges with Heuristics

Our final evaluation presented several challenges in designing credibility nudges. First, participants showed their skepticism towards the selection of heuristics (e.g., bandwagon heuristic) in design. Though the bandwagon heuristic can be robust and hard to falsify, in an open forum such as Twitter, it is open for manipulation. Perhaps, as one of the participants suggested, feedback on the validity of the question may be helpful. Still, problems may also exist with feedback. For one, due to partisan attitude, feedback wars, similar to edit wars in Wikipedia, might result [116]. Additionally, due to frequent updates and the scale of social media content, feedback from volunteers might be scarce on most items. Perhaps a system designer can show the distribution of feedback by users' leanings to reduce the effects of feedback wars and solicit user contributions by promoting items with scarce feedback. Second, participants' concerns with the bandwagon heuristic promoting confirmation bias might be an extension of prior findings that users tend to prefer selecting information (including political information) consistent with their preexisting beliefs [56]. However, scholars have shown that the extent of confirmation bias in information selection, particularly in news selection, is small [45]. In the event of confirmation bias, we can computationally de-bias the heuristic cue, as in prior works [2]. Lastly, audience misperceptions of non-nudged content being credible indicate an additional challenge in design. This effect has also been demonstrated in a recent work [88]. One way to solve this problem would be to add nudges to all content. Aside from these challenges, one participant (U7) pointed out that *"News might change and the user will not see the update when more legitimate questions are added to the replies"*. As user interactions accumulate over time, the number of questions in replies could change, wherein the same tweet would be categorized as *Reliable* at first and *Questionable* at a later time. This change in nudge category stemming from our choice of the bandwagon heuristic could imply an inconsistency in nudge design. System designers can incorporate delayed intervention mechanisms and inform users of this cold-start issue. Overall, these challenges inform designers about considerations for designing nudges with heuristics.

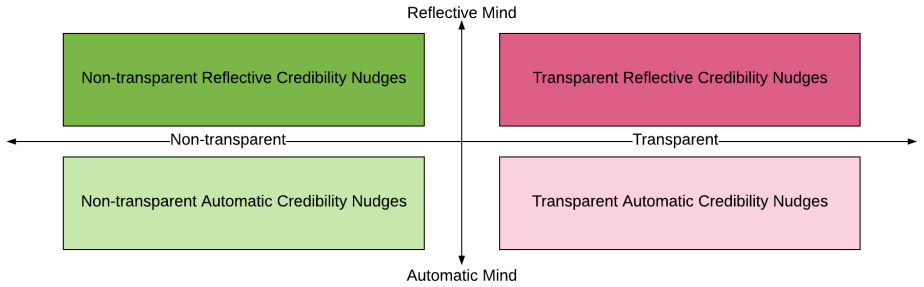


Fig. 6. Types of nudges based on transparency and mode of thinking. This figure emulates Figure 1 by Caraban et al. [21]. This work lies in the bottom-right quadrant.

8 IMPLICATIONS AND OPPORTUNITIES FOR DESIGNING CREDIBILITY NUDGES

We built NudgeCred by fusing the theoretical background on nudges with theories underpinning credibility evaluation. Researchers have the opportunity to explore the nudge categories we built our design around, experiment effectiveness of other heuristics and utilize alternate nudging method. Below, we elaborate a few possibilities around these areas while discussing potential considerations.

Exploring Additional Nudge Categories. Referring to the four categories of nudges, divided along the two dimensions of *transparency* and *mode of thinking* (automatic or reflective), we have illustrated how NudgeCred resides on one of the four quadrants—*transparent automatic* (see Figure 6). Nudge theorists describe transparent nudges with reflective and automatic mode as, respectively, reflective decision making and instinctive decision making [43]. Technology-mediated nudging research focus more on transparent reflective quadrant [21]. For example, Facebook's "Related Articles" feature to tackle misinformation exemplifies such a nudge design [112]. Twitter's blue check "verified" marker on profiles is another example of a transparent automatic credibility nudge. Between reflective and automatic mode of thinking, each has their own benefit. For example, while nudges with automatic mode of thinking can be diminished over time, reflective mode of thinking can educate users and have a lasting impact. On the other hand, design considering reflective mode of thinking requires additional considerations such as motivating the users and assisting in reflection [1]. For example, to motivate users we can show statistics of how often users misread a news tweet as a nudge and to assist them in reflection this nudge can include the list of the most common mistakes. For non-transparent quadrants, scholars propose reflective ones as a manipulation of behavior and automatic ones as a manipulation of choice, in both cases without users' awareness. An example of non-transparent automatic nudging could be showing the average reading time of an article which could prompt users to think shorter articles as less detailed and less credible. Comparatively, research show fewer work in non-transparent reflective quadrant [21]. Due to their deceptive nature, designers need to consider ethical consideration while experimenting on non-transparent quadrants. Overall, this is the start of work in this space; much research needs to be done in cataloging and evaluating news credibility nudges along the other dimensions spanning the four quadrants.

Exploring Alternate Heuristics for Nudging Credibility. Our nudge design is based on two heuristics under the **A**gency affordance, drawn from the MAIN model's list of technology-mediated affordances affecting credibility judgment. Designers have the opportunity to explore the other heuristics in their design. For example, designers could offer design cues to distinguish news items with/without video footage and prompt a *realism heuristic* from **M**odality affordance. Or, design cues distinguishing news tweets with interactive content (e.g., interactive 3D footage or charts) could prompt *interaction heuristic* from **I**nteractivity affordance. For **N**avigability affordance,

designers can prompt *browsing heuristic* by providing additional hyperlinks to journalists' past activities (e.g., MuckRack profile [98]) besides news items.

Examining Alternate Nudging Method. While the original proposers of the concept did not lay out a fixed method for creating successful nudges [108], Caraban et. al. recently devised six broad categories for 23 nudging methods used in prior HCI works, namely, `Facilitate` (e.g., *default choice* and *hiding*), `Confront` (e.g., *remind consequence* and *provide multiple viewpoint*), `Deceive` (e.g., *add inferior alternative* and *deceptive visualization*), `Social Influence` (e.g., *enable comparison* and *public commitment*), `Fear` (e.g., *reduce distance* and *scarcity*) and `Reinforce` (e.g., *ambient feedback* and *subliminal priming*) [21]. Under these categories, NudgeCred utilizes two methods from two categories. First, it works as a `Facilitate` category by facilitating decision making through a combination of color-coding and translucent *hiding* method. Second, it operates as a `social influence nudge` category with *enabling social comparison* method through the use of number of questions asked by other responders. Technology-mediated credibility nudges can utilize other methods from this classification. For example, similar to NewsCube [85], designers can use `confront` category by *offering multiple viewpoint* method on news items. Or, flashing keywords around news items (e.g., "reliable" or "questionable") utilizing *subliminal priming* method under `Reinforce` category [94], could affect users' credibility perception of those items.

Designing Against Adversarial Agents. Our approach to be transparent suggests that adversarial entities, upon knowing algorithmic detail, can manipulate the system. For example, by commenting on a factual mainstream news with a question, they can create a misperception of it being questionable. This is a problem that most online platforms struggle with–balancing between being transparent about their algorithm while safeguarding against adversaries. Indeed, platforms like Reddit while publishing their ranking algorithms include "some fuzzing" to avoid manipulation of their voting system [29]. Hence, some opaqueness in the algorithm might be desirable. At the same time, platform developers could also misuse this fuzziness in the algorithm for their own benefit, such as to drive engagement [44]. Indeed, Twitter's content moderation practice has been controversial in the past, in some cases resulting in reversal of moderation decisions [66]. Similar controversy could arise regarding nudging policy. Therefore, designing nudges requires consideration for the multiple stakeholders of a social platform—platform developer, consumers and news producers. Designers would need to consider the degree to which a nudge would be resistant to adversarial attacks by each stakeholder. For example, crowds' question based bandwagon heuristic has a high-level of susceptibility of manipulation by the consumers compared to the platform developers and news producers. On the other hand, our authority heuristic is more susceptible to manipulation by the platform designers compared to the news producers and consumers. Overall, a potential solution to this problem would be creating a collaborative standard authorized by all stakeholders. For example, Facebook has already created a third-party oversight board for content moderation [13]. A similar strategy can be applied to determine nudging criteria.

Considering Shortcomings of Nudging as a Design Approach. Despite its success, nudging has its own shortcomings. Prior literature proposes that nudges may have diminished effects in the long term for two reasons: (i) nudges relying on automatic cognition may fail to educate the user, and (ii) prolonged exposure may have various effects such as transform nudge effects into background noise [63, 123], invoke a feeling of intrusiveness [133], and reduce users' perception of autonomy [64]. Such effects may lead to unforeseen consequences. For example, a default choice nudge promoting vaccination that reduced users' perception of autonomy resulted in users unsubscribing from the program [64]. In our case, if users repeatedly encounter their favored political news source labeled as *Questionable* or *Unreliable*, they could become averse to the design.

However, designers can apply several strategies to counter this problem. On one end, they can alter the design over time or prompt users to change the intervention settings over time [60]. As an alternate, they can also choose to deploy reflective nudges which are less susceptible to the diminishing effect. A potential problem with the altering design is that news consumers may need to re-learn how nudges operate. Regardless, designers would first have to understand the rate at which nudge effects diminishes, a direction for future research.

9 LIMITATIONS AND FUTURE WORK

Our work is not without limitations. Our Study 1 was conducted in a controlled setting as opposed to in the wild. However, we see two reasons that suggest that our results demonstrating the utility of nudges in credibility assessment could extend to naturalistic Twitter setup as well. First, research shows that self-reported measures pertaining to social media usage correlate with observed behavior in the wild [39, 88]. Second, large-scale survey-based research on nudges pertaining to news on social media, show that nudges affect related attitude, such as sharing intention of misinformation [82, 90]. Our second limitation relates to choice of population. Because we tested variables (e.g., political ideology and media skepticism) that has different meaning across countries, we had to limit our experimental population to the US. To generalize our findings to the global population, future research could replicate our study in the context of each country. Third, our recruitment had limitations that are characteristic of any online service-based recruitment. Though we may not have obtained the true nationally representative sample in the US, research suggests that Qualtrics provides reasonably representative US sample (approximately 6% deviance from the US census) around demography such as gender, age, race and political affiliation [48]. Overall, a large-scale Twitter deployment might reconcile these concerns in the future. We initially attempted to do so by contacting a large number of Twitter users, without much success due to a lack of platform-level cooperation (Our research account was repeatedly blocked). While a 2015 study had successfully piggybacked on Twitter's infrastructure to run large-scale recruitment efforts on the platform [38], we were unable to do so, despite following similar strategy. We anticipate that changes to Twitter's policies may have prevented us from running large-scale recruitment on the platform [51].

10 CONCLUSION

In this study, we provide evidence that a nudge-based design that directs users' attention to specific social cues on the news can affect their credibility judgments. We used three nudges: *Reliable*, applied to mainstream news tweets without questions in replies; *Questionable*, applied to mainstream news tweets with questions in replies; and *Unreliable*, applied to non-mainstream news tweets. Our experiment suggests that users who saw tweets with *Reliable* nudge as more credible, and tweets with *Questionable* and *Unreliable* nudges as less credible compared to the control users. Moreover, our nudges were not affected by users' political preferences, political cynicism, and media skepticism. Through interviews, we found evidence of how nudges can impact users' news consumption and how the current design can be improved. This research proposes further exploration of nudge-based system design approaches for online platforms.

11 ACKNOWLEDGEMENTS

This paper would not be possible without the support from National Science Foundation through grant #2041068. We also appreciate valuable feedback from the members of the Social Computing Lab at Virginia Tech and University of Washington.

REFERENCES

[1] Alexander T Adams, Jean Costa, Malte F Jung, and Tanzeem Choudhury. 2015. Mindless computing: designing technologies to subtly influence behavior. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. 719–730.

[2] Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2014. De-biasing user preference ratings in recommender systems. In *RecSys 2014 Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS 2014)*. Citeseer, 2–9.

[3] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives* 31, 2 (2017), 211–36.

[4] Mike Allen. 1991. Meta-analysis comparing the persuasiveness of one-sided and two-sided messages. *Western Journal of Speech Communication* 55, 4 (1991), 390–404.

[5] AP. 2018. Associated Press News Values and Principles. https://www.ap.org/about/news-values-and-principles/.

[6] Mevan Babakar. 2018. Crowdsourced Factchecking: There is a role for crowdsourcing in factchecking but (so far) it's not factchecking. https://medium.com/@meandvan/crowdsourced-factchecking-4c5168ea5ac3.

[7] Khaled Bachour, Jon Bird, Vaiva Kalnikaite, Yvonne Rogers, Nicolas Villar, and Stefan Kreitmayer. 2012. Fast and frugal shopping challenge. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1459–1460.

[8] Rebecca Balebako, Pedro G Leon, Hazim Almuhimedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Faith Cranor, and Norman Sadeh. 2011. Nudging users towards privacy on mobile devices. In *Proc. CHI 2011 Workshop on Persuasion, Nudge, Influence and Coercion*. Citeseer, 193–201.

[9] Meital Balmas. 2014. When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication research* 41, 3 (2014), 430–454.

[10] Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences* 116, 22 (2019), 10717–10722.

[11] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating Differences in Crowdsourced News Credibility Assessment: Raters, Tasks, and Expert Criteria. *arXiv preprint arXiv:2008.09533* (2020).

[12] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. 2018. FeedReflect: A tool for nudging users to assess news credibility on twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 205–208.

[13] Oversight Board. 2020. Oversight Board | Independent Judgment. Transparency. Legitimacy. https://oversightboard.com/. (Accessed on 04/05/2021).

[14] Ben Bolker. 2018. lme4 package | R Documentation. https://www. rdocumentation.org/packages/lme4/versions/1.1-18-1.

[15] Toby Bolsen and James N Druckman. 2015. Counteracting the politicization of science. *J. Commun.* 65, 5 (2015), 745–769.

[16] Marilynn B Brewer and William D Crano. 2000. Research design and issues of validity. *Handbook of research methods in social and personality psychology* (2000), 3–16.

[17] Campbell Brown. 2019. Introducing Facebook News - About Facebook. https://about.fb.com/news/2019/10/introducing-facebook-news/. (Accessed on 09/11/2020).

[18] Guido Buelow. 2019. Automation Plus Expert Journalism: How Full Fact is Fighting Misinformation - About Facebook. https://about.fb.com/news/2019/06/inside-feed-full-fact-interview/. (Accessed on 09/11/2020).

[19] Julie G Bush, Hollyn M Johnson, and Colleen M Seifert. 1994. The implications of corrections: Then why did you mention it. (1994).

[20] Joseph N Cappella and Kathleen Hall Jamieson. 1997. *Spiral of cynicism: The press and the public good*. Oxford University Press on Demand.

[21] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 503.

[22] D Jasun Carr, Matthew Barnidge, Byung Gu Lee, and Stephanie Jean Tsang. 2014. Cynics and skeptics: Evaluating the credibility of mainstream and citizen journalism. *J. Mass Commun. Q.* 91, 3 (2014), 452–470.

[23] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proc. WWW*. ACM, 675–684.

[24] Mike Caulfield. 2017. Web literacy for student fact-checkers. (2017).

[25] Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, Vol. 5. Hillsdale, NJ: Lawrence Erlbaum, 3–39.

[26] Xinran Chen, Sei-Ching Joanna Sin, Yin-Leng Theng, and Chei Sian Lee. 2015. Deterring the spread of misinformation on social network sites: A social cognitive theory-guided intervention. *Proceedings of the Association for Information*

Science and Technology 52, 1 (2015), 1–4.
[27] Cision. 2017. Top 10 U.S. Daily Newspapers. https://www. cision.com/us/2017/09/top-10-u-s-daily-newspapers-2/.
[28] Russell D Clark III and Anne Maass. 1988. The role of social categorization and perceived source credibility in minority influence. *European Journal of Social Psychology* 18, 5 (1988), 381–394.
[29] Devin Coldewey. 2016. Reddit overhauls upvote algorithm to thwart cheaters and show the site's true scale | TechCrunch. https://techcrunch.com/2016/12/06/reddit-overhauls-upvote-algorithm-to-thwart-cheaters-and-show-the-sites-true-scale/?guccounter=1. (Accessed on 01/03/2021).
[30] John Cook, Stephan Lewandowsky, and Ullrich KH Ecker. 2017. Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS one* 12, 5 (2017), e0175799.
[31] Nicholas Diakopoulos, Munmun De Choudhury, and Mor Naaman. 2012. Finding and assessing social media information sources in the context of journalism. In *Proc. CHI*. ACM, 2451–2460.
[32] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.
[33] Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences* 7, 10 (2003), 454–459.
[34] Facebook. 2019. How Facebook Has Prepared for the 2019 UK General Election - About Facebook. https://about.fb.com/news/2019/11/how-facebook-is-prepared-for-the-2019-uk-general-election/. (Accessed on 09/11/2020).
[35] Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on Twitter?. In *Proceedings of the 2016 ACM SIGMETRICS international conference on measurement and modeling of computer science*. 179–192.
[36] R Kelly Garrett and Brian E Weeks. 2013. The promise and peril of real-time corrections to political misperceptions. In *Proc. CSCW*. ACM, 1047–1058.
[37] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[38] Catherine Grevet and Eric Gilbert. 2015. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proc. CHI*. ACM, 4047–4056.
[39] Andrew Guess, Kevin Munger, Jonathan Nagler, and Joshua Tucker. 2019. How accurate are survey responses on social media and politics? *Political Communication* 36, 2 (2019), 241–258.
[40] Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences* 117, 27 (2020), 15536–15545.
[41] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.
[42] David Halpern. 2016. *Inside the nudge unit: How small changes can make a big difference*. Random House.
[43] Pelle Guldborg Hansen and Andreas Maaløe Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation* 4, 1 (2013), 3–28.
[44] Karen Hao. 2021. He got Facebook hooked on AI. Now he can't fix its misinformation addiction | MIT Technology Review. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/. (Accessed on 04/05/2021).
[45] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin* 135, 4 (2009), 555.
[46] Del Harvey. 2019. Helping you find reliable public health information on Twitter. https://blog.twitter.com/en_us/topics/company/2019/helping-you-find-reliable-public-health-information-on-twitter.html. (Accessed on 09/11/2020).
[47] Del Harvey and David Gasca. 2018. Serving healthy conversation. https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html. (Accessed on 09/11/2020).
[48] MS Heen, Joel D Lieberman, and Terance D Miethe. 2014. A comparison of different online sampling approaches for generating national samples. *Center for Crime and Justice Policy* 1, 9 (2014), 1–8.
[49] John Hegeman. 2018. Facing Facts: Facebook's Fight Against Misinformation - About Facebook. https://about.fb.com/news/2018/05/facing-facts-facebooks-fight-against-misinformation/. (Accessed on 09/11/2020).
[50] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[51] Twitter Inc. 2019. The Twitter Rules. https://help.twitter.com/en/rules-and-policies/twitter-rules. (Accessed on 09/13/2019).

[52] Peter John, Graham Smith, and Gerry Stoker. 2009. Nudge nudge, think think: two strategies for changing civic behaviour. *The Political Quarterly* 80, 3 (2009), 361–370.
[53] Hollyn M Johnson and Colleen M Seifert. 1994. Sources of the continued influence effect: When misinformation in memory affects later inferences. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 6 (1994), 1420.
[54] Shipi Kankane, Carlina DiRusso, and Christen Buckley. 2018. Can we nudge users toward better password management?: An initial study. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, LBW593.
[55] Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 829–839.
[56] Silvia Knobloch-Westerwick and Jingbo Meng. 2009. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research* 36, 3 (2009), 426–448.
[57] Silvia Knobloch-Westerwick, Nikhil Sharma, Derek L Hansen, and Scott Alter. 2005. Impact of popularity indications on readers' selective exposure to online news. *J. Broadcast. Electron. Media* 49, 3 (2005), 296–313.
[58] D. Koehler and N. Harvey. 2004. In D. Koehler & N. Harvey (Eds.). (2004). Blackwell handbook of judgment and decision making (pp. 62–88). Oxford, UK: Blackwell. (2004), 62–88.
[59] Andrew Kohut, Carroll Doherty, Michael Dimock, and Scott Keeter. 2012. Cable leads the pack as campaign news source. *Pew Center for the People and the Press. Retrieved from https://www.pewresearch.org/wp-content/uploads/sites/4/legacy-pdf/2012-Communicating-Release.pdf* (2012).
[60] Geza Kovacs, Zhengxuan Wu, and Michael S Bernstein. 2021. Not Now, Ask Later: Users Weaken Their Behavior Change Regimen Over Time, But Expect To Re-Strengthen It Imminently. *arXiv preprint arXiv:2101.11743* (2021).
[61] Issie Lapowski. 2018. NewsGuard Wants to Fight Fake News with Humans, Not Algorithms.
[62] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045* (2017).
[63] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2011. Mining behavioral economics to design persuasive technology for healthy choices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 325–334.
[64] Birthe A Lehmann, Gretchen B Chapman, Frits ME Franssen, Gerjo Kok, and Robert AC Ruiter. 2016. Changing the default to promote influenza vaccination among health care workers. *Vaccine* 34, 11 (2016), 1389–1392.
[65] Jordan Litman, Tiffany Hutchins, and Ryan Russon. 2005. Epistemic curiosity, feeling-of-knowing, and exploratory behaviour. *Cognition & Emotion* 19, 4 (2005), 559–582.
[66] Elizabeth Lopatto. 2020. In its latest confusing decision, Twitter reinstates The New York Post - The Verge. https://www.theverge.com/2020/10/30/21542801/twitter-lifts-ny-post-ban-policy-changes. (Accessed on 04/05/2021).
[67] Tessa Lyons. 2017. Replacing Disputed Flags With Related Articles - About Facebook. https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/. (Accessed on 09/11/2020).
[68] Tessa Lyons. 2018. Hard Questions: How Is Facebook's Fact-Checking Program Working? - About Facebook. https://about.fb.com/news/2018/06/hard-questions-fact-checking/. (Accessed on 09/11/2020).
[69] Diane M Mackie and Sarah Queller. 2000. The impact of group membership on persuasion: Revisiting "Who says what to whom with what effect?". *Attitudes, behavior, and social context: The role of norms and group membership* (2000), 135–155.
[70] Rakoen Maertens, Jon Roozenbeek, Melisa Basol, and Sander van der Linden. 2020. Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied* (2020).
[71] Zlatina Marinova, Jochen Spangenberg, Denis Teyssou, Symeon Papadopoulos, Nikos Sarris, Alexandre Alaphilippe, and Kalina Bontcheva. 2020. Weverify: Wider and Enhanced Verification for You Project Overview and Tools. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 1–4.
[72] Mediabiasfactcheck. 2019. About - Media Bias/Fact Check. https://mediabiasfactcheck.com/about/. (Accessed on 09/18/2019).
[73] Solomon Messing and Sean J Westwood. 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research* 41, 8 (2014), 1042–1063.
[74] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* 58, 13 (2007), 2078–2091.
[75] Miriam J Metzger, Andrew J Flanagin, and Ryan B Medders. 2010. Social and heuristic approaches to credibility evaluation online. *Journal of communication* 60, 3 (2010), 413–439.
[76] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. 2015. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* (2015), 0093650215613136.

[77] Philip Meyer. 1988. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly* 65, 3 (1988), 567–574.

[78] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. 2014. Political polarization & media habits. http://www.journalism.org/2014/10/21/political-polarization-media-habits/. *Pew Research Center* 21 (2014).

[79] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.

[80] Tanushree Mitra, Graham P Wright, and Eric Gilbert. 2017. A parsimonious language model of social media credibility across disparate events. In *Proc. CSCW*. ACM, 126–145.

[81] Adam Mosseri. 2016. Addressing Hoaxes and Fake News - About Facebook. https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/. (Accessed on 09/11/2020).

[82] Elmie Nekmat. 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 (2020), 2056305119897322.

[83] Onora O'neill. 2002. *A question of trust: The BBC Reith Lectures 2002*. Cambridge University Press.

[84] OpenSources. 2018. OpenSources. http://opensources.co.

[85] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 443–452.

[86] Jessica Paynter, Sarah Luskin-Saxby, Deb Keen, Kathryn Fordyce, Grace Frost, Christine Imms, Scott Miller, David Trembath, Madonna Tucker, and Ullrich Ecker. 2019. Evaluation of a template for countering misinformation—Real-world Autism treatment myth debunking. *PloS one* 14, 1 (2019).

[87] Cornelia Pechmann. 1992. Predicting when two-sided ads will be more effective than one-sided ads: The role of correlational and correspondent inferences. *Journal of Marketing Research* 29, 4 (1992), 441–453.

[88] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* (2020).

[89] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David Rand. 2019. Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv. com/3n9u8* (2019).

[90] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.

[91] Gordon Pennycook and David G Rand. 2017. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. (2017).

[92] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.

[93] Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*. Springer, 1–24.

[94] Charlie Pinder, Jo Vermeulen, Russell Beale, and Robert Hendley. 2015. Subliminal priming of nonconscious goals on smartphones. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. 825–830.

[95] Sara Pluviano, Caroline Watt, and Sergio Della Sala. 2017. Misinformation lingers in memory: failure of three pro-vaccination strategies. *PLoS One* 12, 7 (2017), e0181640.

[96] THE MEDIA INSIGHT PROJECT. 2018. Americans and the News Media: What they do — and don't — understand about each other - American Press Institute. https://www.americanpressinstitute.org/publications/reports/survey-research/americans-and-the-news-media/. (Accessed on 09/07/2020).

[97] Vahed Qazvinian, Emily Rosengren, Dragomir R Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1589–1599.

[98] Muck Rack. 2019. Muck Rack | For Journalists. https://muckrack.com/journalists. (Accessed on 09/14/2019).

[99] Marilyn Roberts, Wayne Wanta, and Tzong-Horng Dzwo. 2002. Agenda setting and issue salience online. *Communication research* 29, 4 (2002), 452–465.

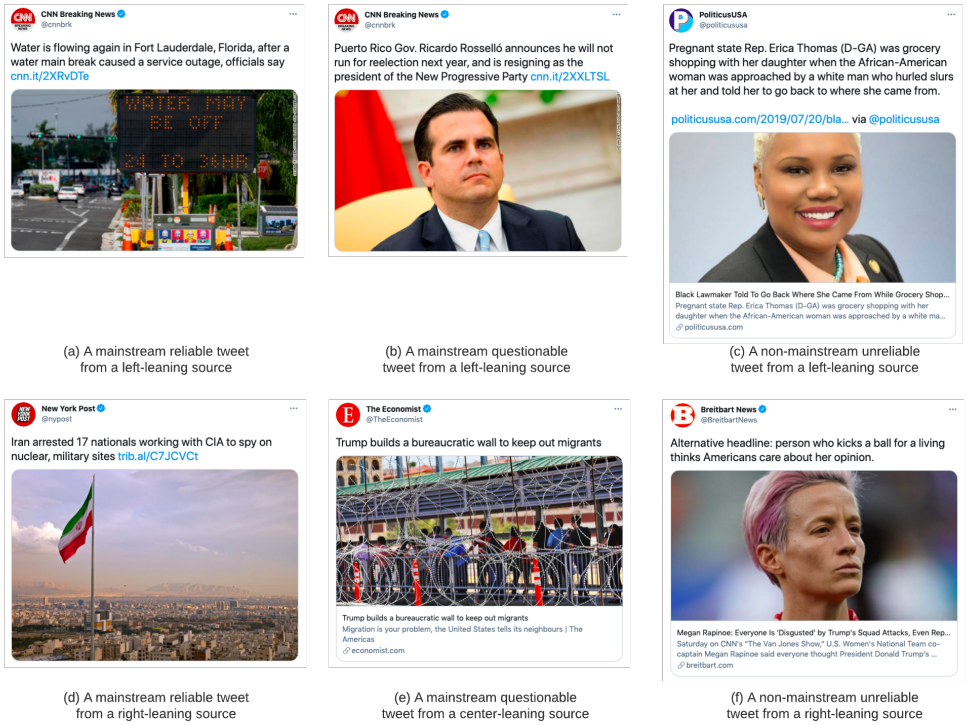
[100] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 1–10.

[101] Guy Rosen. 2020. An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19 - About Facebook. https://about.fb.com/news/2020/04/covid-19-misinfo-update/. (Accessed on 09/11/2020).

[102] Guy Rosen and Tessa Lyons. 2019. Remove, Reduce, Inform: New Steps to Manage Problematic Content - About Facebook. https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/. (Accessed on 09/11/2020).

[103] Yoel Roth and Ashita Achuthan. 2020. Building rules in public: Our approach to synthetic & manipulated media. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html. (Accessed on 09/11/2020).
[104] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html. (Accessed on 09/11/2020).
[105] Twitter Safety. 2019. Strengthening our approach to deliberate attempts to mislead voters. https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-approach-to-deliberate-attempts-to-mislead-vot.html. (Accessed on 09/11/2020).
[106] Twitter Safety. 2020. Expanding our policies to further protect the civic conversation. https://blog.twitter.com/en_us/topics/company/2020/civic-integrity-policy-update.html. (Accessed on 09/11/2020).
[107] Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz, and Claire Wardle. 2020. It matters how platforms label manipulated media. Here are 12 principles designers should follow. - The Partnership on AI. https://www.partnershiponai.org/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow/. (Accessed on 07/01/2020).
[108] Evan Selinger and Kyle Whyte. 2011. Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass* 5, 10 (2011), 923–935.
[109] Alina Selyukh, Maria Hollenhorst, and Katie Park. 2016. Disney-Fox Deal: Who Controls Digital Media? Conglomerates, Brands In One Chart|NPR. https://www.npr.org/sections/alltechconsidered/2016/10/28/ 499495517/big-media-companies-and-their-many-brands-in-one-chart.
[110] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2018. The spread of low-credibility content by social bots. *Nature communications* 9, 1 (2018), 1–9.
[111] Henry Silverman. 2019. Helping Fact-Checkers Identify False Claims Faster - About Facebook. https://about.fb.com/news/2019/12/helping-fact-checkers/. (Accessed on 09/11/2020).
[112] Jeff Smith. 2019. Designing Against Misinformation - Facebook Design - Medium. https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2. (Accessed on 09/18/2019).
[113] Jeff Smith, Alex Leavitt, and Grace Jackson. 2018. Designing New Ways to Give Context to News Stories | Facebook Newsroom. https://newsroom.fb.com/news/2018/04/inside-feed-article-context/.
[114] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17 (1994), 273–85.
[115] Jay Sullivan. 2020. Introducing a Forwarding Limit on Messenger - About Facebook. https://about.fb.com/news/2020/09/introducing-a-forwarding-limit-on-messenger/. (Accessed on 09/11/2020).
[116] Róbert Sumi, Taha Yasseri, et al. 2011. Edit wars in Wikipedia. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 724–727.
[117] S Shyam Sundar. 2008. The MAIN model: A heuristic approach to understanding technology effects on credibility. *Digital media, youth, and credibility* 73100 (2008).
[118] S Shyam Sundar, Haiyan Jia, T Franklin Waddell, and Yan Huang. 2015. Toward a theory of interactive media effects (TIME). *The handbook of the psychology of communication technology* (2015), 47–86.
[119] S Shyam Sundar, Anne Oeldorf-Hirsch, and Qian Xu. 2008. The bandwagon effect of collaborative filtering technology. In *CHI'08 extended abstracts*. ACM, 3453–3458.
[120] Cass R Sunstein. 2014. *Why nudge?: The politics of libertarian paternalism*. Yale University Press.
[121] Cass R Sunstein. 2015. The ethics of nudging. *Yale J. on Reg.* 32 (2015), 413.
[122] Cass R Sunstein. 2016. Do people like nudges? (2016).
[123] Cass R Sunstein. 2017. Nudges that fail. *Behavioural Public Policy* 1, 1 (2017), 4–25.
[124] Twitter Support. 2020. New labels for government and state-affiliated media accounts. https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html. (Accessed on 09/11/2020).
[125] David Tannenbaum, Craig R Fox, and Todd Rogers. 2017. On the misplaced politics of behavioural policy interventions. *Nature Human Behaviour* 1, 7 (2017), 0130.
[126] Richard H Thaler. 2008. Nudge: Improving decisions about health, wealth, and happiness.
[127] Sreethu Thulasi. 2019. Understand Why You're Seeing Certain Ads and How You Can Adjust Your Ad Experience - About Facebook. https://about.fb.com/news/2019/07/understand-why-youre-seeing-ads/. (Accessed on 09/11/2020).
[128] Peter M. Todd and Gerd Gigerenzer. 2000. Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences* 23, 5 (2000), 727–780. https://doi.org/10.1017/S0140525X00003447
[129] Yariv Tsfati. 2010. Online news exposure and trust in the mainstream media: Exploring possible associations. *American Behavioral Scientist* 54, 1 (2010), 22–42.
[130] Yariv Tsfati and Joseph N Cappella. 2003. Do people watch what they do not trust? Exploring the association between news media skepticism and exposure. *Communication Research* 30, 5 (2003), 504–529.

[131] Yariv Tsfati and Joseph N Cappella. 2005. Why do people watch news they do not trust? The need for cognition as a moderator in the association between news media skepticism and exposure. *Media psychology* 7, 3 (2005), 251–271.
[132] Twitter. 2019. Notices on Twitter and what they mean. https://help.twitter.com/en/rules-and-policies/notices-on-twitter. (Accessed on 09/11/2020).
[133] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. 2014. A field trial of privacy nudges for facebook. In *Proc. CHI*. ACM, 2367–2376.
[134] Yang Wang, Pedro Giovanni Leon, Kevin Scott, Xiaoxuan Chen, Alessandro Acquisti, and Lorrie Faith Cranor. 2013. Privacy nudges for social media: an exploratory Facebook study. In *Proceedings of the 22nd International Conference on World Wide Web*. 763–770.
[135] Claire Wardle. 2019. Misinformation Has Created a New World Disorder - Scientific American. https://www.scientificamerican.com/article/misinformation-has-created-a-new-world-disorder/. (Accessed on 05/27/2020).
[136] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41, 1 (2019), 135–163.
[137] Robin Worrall. 2020. Social media used to spread, create COVID-19 falsehoods – Harvard Gazette. https://news.harvard.edu/gazette/story/2020/05/social-media-used-to-spread-create-covid-19-falsehoods/. (Accessed on 09/12/2020).
[138] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604* (2016).
[139] Amy X Zhang, Aditya Ranganathan, Sarah Emlen Metz, Scott Appling, Connie Moon Sehat, Norman Gilmore, Nick B Adams, Emmanuel Vincent, Jennifer Lee, Martin Robbins, et al. 2018. A structured response to misinformation: Defining and annotating credibility indicators in news articles. In *Companion Proceedings of the The Web Conference 2018*. 603–612.
[140] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*. 1395–1405.



(a) A mainstream reliable tweet from a left-leaning source

(b) A mainstream questionable tweet from a left-leaning source

(c) A non-mainstream unreliable tweet from a left-leaning source

(d) A mainstream reliable tweet from a right-leaning source

(e) A mainstream questionable tweet from a center-leaning source

(f) A non-mainstream unreliable tweet from a right-leaning source

Fig. 7. Sample tweets used in Study 1 without the interventions. The examples include reliable, questionable and unreliable tweets from left-/center-/right-leaning sources. Here, there is a mix of politically contentious (e.g., immigration, racism and LGBTQ+) and not so contentious issues (e.g., flood and national security).

A EXAMPLE TWEETS USED IN STUDY 1

Figure 7 shows several tweets used in Study 1. These tweets were selected by finding the most popular tweets from the last 48 hours from various partisan sources. Notice the partisan nature of the items including immigration, racism and LGBTQ+ issues.

B STUDY 2: INTERVIEW QUESTIONNAIRE

- Could you tell me about your news reading on twitter? How often do you read news? and what type of news do you read?
- How often do you come across misinformation?
- Have you ever felt the need for any tools to improve news reading on twitter?
- Have you ever used any tools to improve news reading on twitter? What tool? How did it work?
- (Asking them to share their screen for twitter feed) How would you compare your Twitter use during the study to how you normally do?
- Could you tell me about a time when you paid more attention to a news on Twitter in the last 5 days? Why? What was it about?
- How satisfied were you with what you saw on Twitter in the last 5 days?
- What aspect of the intervention did you notice most? How did that impact you?

- When using the extension, did you think of anything it was missing? What more should it do?
- Can you think of any other application of this extension that you would like?
- During the study, did you have any issues with NudgeCred? Is there any aspect of the usability (eg., design, speed, accuracy) you thought could be improved?
- How did you feel about the plugin overall? What did you like about it? What did you dislike about it?
- Would you continue using it after this study?

Received January 2021; revised April 2021; accepted July 2021