Appealing to Sense and Sensibility: System 1 and System 2 Interventions for Fake News on Social Media

Patricia L. Moravec
Information, Risk, and Operations Management Department, McCombs School of Business,
University of Texas at Austin, Austin TX 78705
patricia.moravec@mccombs.utexas.edu

Antino Kim
Operations and Decision Technologies Department, Kelley School of Business,
Indiana University, Bloomington IN 47405
antino@indiana.edu

Alan R. Dennis
Operations and Decision Technologies Department, Kelley School of Business,
Indiana University, Bloomington IN 47405
ardennis@indiana.edu

ABSTRACT

Disinformation on social media—commonly called "fake news"—has become a major concern around the world, and many fact-checking initiatives have been launched to mitigate the problem. The way fact-checking results are presented to social media users is important because if the presentation format is not persuasive, fact checking may not be effective. For instance, Facebook tested the idea of flagging dubious articles in 2017 but concluded that it was ineffective and subsequently removed the feature. We conducted three experiments with social media users to investigate two different approaches to implementing a fake news flag, one designed to have its primary effect when processed by automatic cognition (System 1) and the other designed to have its primary effect when processed by deliberate cognition (System 2). We found that both interventions were effective, and an intervention that combined both approaches was about twice as effective. We also found that awareness training on the meaning of the flags increased the effectiveness of the System 2 intervention, but not the System 1 intervention, exactly as theory predicts. Believability, in turn, influenced the extent to which users would engage with the article (e.g., read, like, comment, and share). Our results suggest that both theoretical routes can be used—separately or together—in the presentation of fact-checking results in order to reduce the influence of fake news on the users.

Keywords: Social media, fake news, disinformation, dual process cognition, credibility.

INTRODUCTION

Fake news has been defined as "news articles that are intentionally and verifiably false and could mislead readers" (Allcott and Gentzkow 2017), and its spread on social media has had important effects on the users, as well as on the society as a whole (Shane 2017; Wakabayashi and Shane 2017). More than 60% of adults read news on social media (Gottfried and Shearer 2016), and the figure is even higher among college students (Head et al. 2018). Fake news has become a major societal concern because, beyond simply misleading readers on a single issue, it creates an environment in which people do not know what to believe (Barthel et al. 2016; Head et al. 2018). Social media users are unable to detect truth from fiction (Barthel et al. 2016; Kim and Dennis 2019; Moravec et al. 2019) possibly because of *confirmation bias*: users are more likely to believe news articles—real or fake—that align with their prior beliefs (Kim and Dennis 2018). Many fact-checking initiatives have been launched to address users' inability to detect fake news (Graves 2016).

Fake news is a major problem on social media such as Facebook (Shane 2017)—the most ubiquitous social media platform in the world, with more than 2 billion users (Barthel et al. 2016; Statista 2018)—partly because Facebook is unlike other sources of news. On Facebook, users do not choose the sources of the articles that they see. Articles from many different sources are intermixed, which is unlike other technologies where users explicitly choose the source (e.g., TV news, news websites, news apps on mobile phones). On Facebook, articles that our friends share appear together with articles from past sources we have read and articles from advertisers that have paid to place them in our news feed. These articles may be true or fake, designed to deliberately influence opinions (Shane 2017), whether they appear on our newsfeeds from advertisers or from friends who have accidentally or intentionally shared them; 23% of social media users report that they have spread fake news, either accidentally or intentionally (Barthel et al. 2016). That statistic only refers to those who are aware that they have spread fake news. Unfortunately, this is likely a lower bound, as it has been found that people, not bots, tend to be the primary diffusers of false

stories, which actually spread faster than true stories (Vosoughi et al. 2018). On social media, anyone can create "news"—real or fake—and the news spreads throughout the Internet as social media users read it and share it (Mohammed 2012). Quality control has moved from journalists with a putative interest in truth to users who have no training and often do not verify facts before spreading news (Kim and Dennis 2019).

Fake news on social media is also a problem because of the way users consume information on social media. Fake information has been a problem in other Internet contexts as well, such as product reviews (Cheung et al. 2012; Lappas et al. 2016) or health information (Bernstam et al. 2005; Eysenbach et al. 2000). The key difference between reading social media and reading product reviews or health information is the user's mindset (Ho et al. 2017; Li and Hitt 2008; Yin et al. 2016). Most people use social media for pleasure (Johnson and Kaye 2015; Sledgianowski and Kulviwat 2009), which means that they are in a hedonic mindset, as opposed to a utilitarian one (Harsanyi 1977). In other words, people use social media to detach from work and unwind—to "deactivate" (Panger 2018). In contrast, most people reading product reviews or health information are striving to make a decision and thus are in a utilitarian mindset (Cotte et al. 2006); few people read product reviews or health information for entertainment. Individuals in a hedonic mindset are less likely to critically consider information than those in a utilitarian mindset, and thus, they are less likely to engage in deep consideration of the information they see (Kahneman 2003).

In an effort to alert users of misleading articles, in 2017, Facebook introduced a "disputed" flag which was placed on fake news articles. However, Facebook withdrew this flag in 2018 after determining that it was ineffective and may have even backfired (Meixler 2017). There are several potential reasons why the flag did not work as intended, such as motivating users to click on the potentially fake article out of curiosity or influencing them to believe it more (Lyons 2017). Based on a mental model by Bravo-Lillo et al. (2010), when a warning message pops up, users may (i) observe and consider, (ii) judge and decide, and (iii) act. The reason for the failure of Facebook's flag may lie in the first stage of this mental model, or second, or both. In this paper, we propose two theoretical explanations for the flag's shortfall and, accordingly, two potential improvements to increase the effectiveness of fact checking.

The first is knowledge. It is possible that social media users did not fully comprehend the meaning

of Facebook's flag, and thus it had little influence over the users (Anderson and Agarwal 2010). Worse, the flag may act as an attention-grabber and make the users believe the flagged articles even more. Hence, actively informing the users so that they are aware that the flag indicates a fake news article may prove useful. This should be particularly effective in the flag's introductory stage since a more formal education would likely enable quicker understanding of the flag's meaning. Overtime, as users become more informed about the flag through daily usage and informal channels, formal education may become unnecessary.

The second possible reason is firmly rooted in cognitive psychology. Humans have two separate cognition processes: System 1 cognition is automatic and produces instant judgments in less than a second, while System 2 cognition is deliberate and requires cognitive effort (Kahneman 2011; Stanovich and West 2000). Individuals usually do not invoke System 2 cognition unless they notice something amiss that requires deeper thought (Kahneman 2011). Because of a hedonic mindset, social media users' consumption of social media news is likely to be less mindful (Thatcher et al. 2018). Thus, possible approaches to improving Facebook's flag would be to change its design to be more effective when processed by the user's System 1 cognition, System 2 cognition, or both. In other words, an effective flag would be one that (i) generates a strong gut reaction to stop users from passively consuming information and (ii) persuades the users that the flagged articles are fake.

In this paper, we investigate the extent to which training and flags designed to be more effective under System 1 cognition, System 2 cognition, or both, affect social media users' beliefs of news articles, as well as their likelihood of engaging with them (i.e., read, like, comment, and share). We found that training significantly influenced the effectiveness of different flags on users' beliefs. Thus, Facebook could have made its flags more effective by educating the users. Similarly, we found that the flag designed to be most effective when processed by System 1 cognition and the flag designed to be most effective for System 2 cognition both had significant effects. Training increased the effectiveness of the flag designed for System 2 cognition but as theorized, the flag designed for System 1 was unaffected by training because training should have no effect on System 1's instant intuitive cognition (Dennis and Minas 2018). The flag that combined both System 1 and System 2 interventions became more than twice as strong after training.

Therefore, we argue that an effective fake news flag would be one that is designed for both System 1 and System 2 cognition, and one that users have become familiarized with through awareness training.

PRIOR THEORY AND RESEARCH

When considering information consumption, context matters (Johns 2006; Johns 2017). Prior work in IS has examined work contexts, but our focus is on social media, where individuals primarily browse for hedonic purposes (Chauhan and Pillai 2013). Hence, in the context of social media, fake news may have greater influence, as individuals are not in a mindset to expend effort to detect and respond to false claims.

Fake news has become a part of social media and one of the top social issues after the 2016 U.S. presidential election (Allcott and Gentzkow 2017; Barthel et al. 2016). Many political fake news articles shared throughout the election originally came from various sensationalist American sites or sites controlled by Russia, but once repurposed by teenagers in Macedonia, found new traction on social media (Kirby 2016). By creating news outlets with reasonable names, the teenagers made money through users' clicks on their fake articles (Kirby 2016). Surprisingly, more fake news articles were shared than real news on social media during these months (Silverman 2016), and, unfortunately, people tended to believe the false articles (Barthel et al. 2016; Silverman and Singer-Vine 2016). Some argue that the outcome of the U.S. Presidential election was influenced by fake news (Allcott and Gentzkow 2017; Parkinson 2016).

More than 16% of Americans report that they have unwittingly shared fake news, with 64% saying that fake news leaves them with a considerable amount of confusion about the basic facts of current events (Barthel et al. 2016). Many organizations have launched fact-checking initiatives that attempt to combat fake news on social media (Graves 2016; Lowrey 2017). Facebook's initial response to fake news was to deny any wrong-doing, but they eventually recognized their role in the rapid diffusion of fake news throughout 2016 and slowly began to take actions to combat the spread of fake news (Isaac 2016).

Fact Checking and Facebook's Flag

In response to fake news, a number of fact-checking initiatives were launched (Graves 2016; Lowrey 2017). Some are older initiatives that existed long before the fake news problem on social media was widely recognized. For example, Politifact has been checking specific statements made by politicians

and has been rating the statements for accuracy since 2007. The efforts of Politifact were intended to reduce the opacity of politicians' statements to enable people to better govern themselves.

There have also been several attempts to automate fact checking, such as *Truthy* (Ratkiewicz et al. 2011) and *Hoaxy* (Shao et al. 2016), so that the results can be provided more quickly. *Hoaxy,* probably the best-known technical solution, searches fact-checking sites that verify articles and sites with a history of publishing fake news to build a database of articles. It routinely monitors the spread of articles in real time (e.g., by monitoring Twitter) and displays both the spread of articles and the results of fact checking.

Fact checking has been shown to influence the credibility of political candidates (Wintersieck 2017). Wintersieck (2017) showed participants a video of New Jersey gubernatorial debates, varying the results of fact checks for the two candidates. The fact-checking results changed the perceived debate performance of the candidates and influenced the intention to vote for different candidates.

But fact checking can only work as long as users are aware of it, and one issue with fact checking is that it is usually presented in a different place than the article itself. Thus, users have to be motivated to seek out the fact checking site and find its assessment of the article on social media. There are obvious advantages to including the results of fact checking with the article itself.

In December, 2016, Facebook developed an approach that integrated the results of fact-checking into the display of social media articles (Facebook Help Center 2017). If enough users reported an article as suspicious, then the article was passed to third-party fact checkers for assessment. If these fact checkers determined the article to be fake, Facebook continued to show the article but appended a fake news warning flag stating that the article was "disputed by 3rd party fact-checkers" (see Figure 6b). Facebook discontinued the flag in late 2017 because it determined that the flag was not effective (Meixler 2017).

Facebook's flag followed the common norm for warning messages (Anderson et al. 2016; Vance et al. 2018) in that it had both a graphical icon and a text message. Warning messages are typically designed this way because it has long been theorized that the message needs to first grab the user's attention (hence an icon) and then explain the meaning such that the user comprehends the situation and becomes motivated to engage in a certain behavior (hence the text) (Cranor 2008; Wogalter et al. 2002). A graphical icon is

often used to draw the user's attention, thus making the message more salient (Anderson et al. 2016; Kelley et al. 1989; Wogalter et al. 2002), with the wording of the accompanying text used to explain the situation and desired behavior (Wogalter et al. 2002). The graphical icon can also be used to augment the text in communicating desired behavior (Kelley et al. 1989; Wogalter et al. 2002).

Research shows that warning messages can affect behavior. For instance, warning labels and messages can promote healthy and responsible behavior (Auer and Griffiths 2015; Bansal-Travers et al. 2011; Wohl et al. 2013). Unlike research on interface design, which tends to focus on instantiation (e.g., size, color, positioning, specific wording) (Bansal-Travers et al. 2011), we focus on underlying theory that would ultimately guide design principles. Indeed, designing warning messages using theory is considered a best practice from the perspective of interface design (Wogalter et al. 2002; Wogalter et al. 1999).

Awareness Training

One possible reason for the failure of Facebook's flag could be that the users simply did not understand its meaning. If users did not understand that it signaled fake articles, they may have ignored it or paid more attention to the flagged articles because the flag grabbed their attention.

One way to test this theoretical argument would be to provide users with brief awareness training on the meaning of the flag. If awareness training makes the flag more effective, then this would suggest that its ineffectiveness was due to a lack of understanding. The training that we describe here—and conduct in our experiment—is similar to the awareness training often offered by Facebook; it is similar in nature to *Security Education Training and Awareness* (SETA) programs in the context of cybersecurity (D'Arcy et al. 2009; Johnston et al. 2015; Straub Jr 1990). Rather than expending excessive effort in training users (e.g., educating users and testing their abilities afterwards), awareness training is presented in a short paragraph or a video clip.

Thus, awareness training for a fake news flag can be easily provided on social media. Facebook often displays alerts at the top of users' walls to push important information; see Figure 1. As users understand the meaning of features better, we can expect those features to have a stronger influence. In summary, awareness training on the meaning of a fake news flag will increase its effects. Users will be less

likely to believe an article with a fake news flag after they receive training on the meaning of the flag. Thus:

H1: *Training on the meaning of a fake news flag will increase its negative effect on the believability of news articles on social media.*

[Insert Figure 1]

Cognition in Social Media

We often assume humans are rational actors making deliberate decisions, but research in psychology suggests that this is often not the case (Kahneman 2011). Psychology research has long argued that there are two fundamentally different types of cognition. Automatic System 1 cognition occurs first because it runs continuously and involuntarily provides conclusions without conscious thought (Kahneman 2011). System 2 cognition, which may occur second, involves more effortful, deliberate cognition (Kahneman 2011). Both can influence how we use information technologies (Khatri et al. 2018).

System 1 and System 2 Cognition. System 1 cognition is our "fast thinking," where our simple heuristics produce perceptions and actions in less than a second (Kahneman 2011). System 1 is intuitive (Achtziger and Alós-Ferrer 2013); it is our "gut reaction". The quick nature of System 1 is what enables us to do intuitive tasks without direct thought or attention paid to the task, such as walk, talk, recognize faces, and effortlessly retrieve certain facts from memory (Kahneman 2011). However, System 1 comes with certain drawbacks. When we process information using System 1 cognition, we only use the information immediately at hand; the vividness and saliency of that information drives decisions, rather than a more nuanced and carefully considered approach (de Castro Bellini-Leite 2013; Kahneman 2011). As long as we can form the information into a coherent story—right or wrong—we feel comfortable with our immediate System 1 response (Dennis and Minas 2018). We cannot avoid System 1 cognition because it runs continuously without voluntary control and is constantly telling us what to do (Evans and Stanovich 2013; Kahneman 2011; Thompson 2013). Likewise, it is difficult for us to censor our System 1 response, as its answer and the story that we have developed to support it feel intuitive and credible (Kahneman 2011).

System 2 is our deliberate cognition system, which takes much more time to arrive at a conclusion. This "slow thinking" is laborious (Kahneman 2011; Loewenstein et al. 2015). There are physiological

symptoms that indicate the effort involved: our pupils dilate, heart rate increases, blood pressure rises, and extra blood flows to the different active areas of the brain (Kahneman 2011). Examples include doing mental math (i.e., holding numbers in working memory), monitoring our behavior in tense social situations, comparing two products for value, and checking the validity of a complex argument (Mograbi 2011).

Humans are predisposed to avoid System 2 unless there is a need for it because it is effortful (Dennis and Minas 2018; Kahneman 2011; Stanovich and West 2000). We usually adopt the perceptions and actions produced by System 1 unless we are motivated to invest effort or our System 1 warns us that something is amiss (Bago and De Neys 2017; Kahneman 2011). If we are not willing to expend cognitive effort, we are more likely to believe new information that aligns with our past beliefs due to the gullible nature of our System 1. This *confirmation bias* is driven by the associative memory processing of System 1, which quickly searches for confirming evidence of the question posed. The questions "Is Pat friendly?" and "Is Pat unfriendly?" are fundamentally different questions because they trigger our System 1 to retrieve entirely different instances of Pat's behavior depending on which question we hear (Kahneman 2011).

We can override the results of System 1 cognition by using the deliberate cognition of System 2 (Kahneman 2011). The System 1 results are integrated into our mental model and stored in working memory alongside the facts of the situation (Srull and Wyer 1980; Thompson 2013). Thus, the results produced by System 1 influence any System 2 cognition that follows (Dennis and Minas 2018). The net result is that System 1 cognition (and the confirmation bias it promotes) is a powerful force in perception and behavior (de Guinea and Markus 2009; Dennis and Minas 2018; Kahneman 2011). Although users may invoke System 2 cognition, their decisions and feelings toward news—true or fake—are influenced by System 1.

User's Mindset on Social Media. Our predisposition to use System 1 cognition is exacerbated by social media as most users are in a hedonic mindset (Harsanyi 1977) when they use social media (Johnson and Kaye 2015; Sledgianowski and Kulviwat 2009). Social cognition (i.e., cognition regarding humans and human affairs) is inversely correlated with a utilitarian mindset (Greene et al. 2008). People generally use social media for hedonic purposes such as entertainment, connecting with friends, watching interesting videos, seeking jokes, and so on (Johnson and Kaye 2015; Sledgianowski and Kulviwat 2009). Research

suggests that those in a hedonic mindset are less likely to critically consider information (Cotte et al. 2006; Hirschman and Holbrook 1982) because they are reluctant to spend the effort required for System 2. Instead, they base their opinions on the intuitive judgements produced by System 1 (Kahneman 2003).

The hedonic mindset of social media users drives their interactions with articles as well as their information processing. Simply put, users do not engage deeply with what they read. A study of Twitter click-stream data showed that 59% of article retweets (the primary form of sharing on Twitter) were done *without* the user clicking on the article link (Gabielkov et al. 2016). In other words, 59% of times in which someone decided to share an article, the decision was made *without* reading it. Moreover, 55% of those who clicked on an article link left the page in less than 15 seconds (Haile 2014). This does not give adequate time for System 2 to process the entire article's information and make an informed decision on its veracity. Combining these two statistics leads us to deduce that more than 80% of articles that are shared on social media are likely processed by System 1 cognition. Even in the hopeful case where this figure is an overestimation, there is still a meaningful number of social media users sharing information without using their System 2 cognition to critically consider the information they are sharing.

System 1 cognition is strongly influenced by confirmation bias (Bago and De Neys 2017; Kahneman 2011) because it draws heavily on past experience; it prefers information that matches existing beliefs (Nickerson 1998). Confirmation bias is marked by a prejudice against information that challenges prior beliefs, which results in a disregard toward opposing facts (Nickerson 1998). This disregard leads users to ignore information that does not fit their pre-conceived notions (Devine et al. 1990; Koriat et al. 1980; Moravec et al. 2019).

People are likely to have an opinion about news (Knobloch-Westerwick and Lavis 2017), which comes in the form of confirmation bias (Ask and Granhag 2005; Nickerson 1998; Park et al. 2013). This bias is difficult to counteract as most users are not aware of their own biases, and even those who are aware are unable to recognize each judgment they make (Kahneman 2011). While tasks carried out in utilitarian settings are also susceptible to confirmation bias (Kahneman 2011), hedonic settings exacerbate it because people are less likely to exert cognitive effort (Cotte et al. 2006; Hirschman and Holbrook 1982). Since

users "deactivate" on social media—e.g., leisurely going through friends' vacation photos and watching funny video clips—it is difficult to motivate them to think critically when they suddenly encounter news articles within the same stream of posts. Making the matter worse, social media deliberately display articles conforming to user's preference (The Wall Street Journal 2016). This results in a decrease in the variability of information shown and what has become known as an echo chamber—where the user continually sees opinions that echo his or her own, reinforcing the user's preexisting beliefs (Cerf 2016; The Wall Street Journal 2016). Past research shows that confirmation bias is a major factor affecting social media user's belief in the articles they see (Kim and Dennis 2019; Moravec et al. 2019).

In summary, we theorize that social media users predominantly use System 1 cognition which leads to confirmation bias: they believe articles that align with prior beliefs and disregard those that do not. Our focus is fake news, so this bias is mainly an issue when the social media article is false but aligns with the user's beliefs; in this case, users are likely to believe the fake article. One solution is to cause a strong gut-level reaction (i.e., System 1) to stop believing the information. Another is to trigger cognitive dissonance because strong cognitive dissonance often invokes System 2 cognition to resolve it (Kahneman 2011).

Cognitive Dissonance. Cognitive dissonance occurs when users are presented with two pieces of information that both cannot be true (Festinger 1962; Mills 1999). This contradiction causes cognitive discomfort (Aronson 1969), and users either ignore the discomfort or invoke System 2 to resolve it. If the cognitive dissonance is weak or the issue is unimportant to the users, they often ignore it. If it is strong, users may invoke System 2 cognition to think critically (Aronson 1969).

There are multiple ways to trigger the need to resolve cognitive dissonance (Mills 1999). Facebook developed a fake news flag that was appended to articles that third-party fact-checkers had determined to be false (Levin 2017). Ideally, a flag would trigger the user to either disregard the headline initially with System 1 cognition, or pause, invoke System 2 cognition, and not believe the article. Of course, if the user already thought the article was false, the flag would have no additional value. Unfortunately, Facebook concluded that the flag was not effective and abandoned it (Anderson and Agarwal 2010; Meixler 2017).

There are two possible theoretical interpretations of these events. One is that no fake news flag can

overcome confirmation bias; everyone is already firmly set in their beliefs, and there is no flag that can make the users become skeptical of fake news articles that they agree with. The other possibility is that the flag designed by Facebook was not strong enough to overcome confirmation bias. We argue that a test of one instantiation of a fake news flag by an admittedly reluctant corporation concerned about the flag's effects on users' engagement with content is not sufficient to warrant the first conclusion. Therefore, it remains an open question whether or not a carefully designed fake news flag can nudge users to not believe fake news, either by influencing their gut reaction (i.e., System 1) or by creating a strong cognitive dissonance to invoke critical thinking (i.e., System 2). Hence, we explore two different routes to improve the flag that Facebook may have prematurely abandoned.

Interventions Designed for System 1 and System 2

There are two distinct types of cognition, so there are two different approaches we can take to design a technology intervention intended to induce strong cognitive dissonance: one designed to act *primarily* through System 1 cognition and one designed to act *primarily* through System 2 cognition. The paths are not separate and distinct because System 2 follows System 1. Thus, any intervention that is seen *will* be processed by System 1, and any intervention *may* be processed by System 2.

Thus, while fully recognizing that there is a connection between System 1 and System 2, we theorize that there are two different ways *to approach the design of an intervention*. For simplicity, we will use the terms *System 1 intervention* and *System 2 intervention*, but it is important to note that these terms refer to interventions designed to act *primarily—not* exclusively—through one type of cognition or the other; see Figure 2. The System 1 intervention may influence subsequent System 2 cognition, but its first-order *intended* effect is to influence the intuitive System 1 cognition; thus, a well-designed System 1 intervention will have its primary effect when processed by System 1 and additional System 2 processing will have little additional benefit. Similarly, the System 2 intervention must be processed first by System 1 cognition, but the intervention is *intended* to primarily influence perceptions and behavior when it is processed by deliberate and effortful System 2 cognition; the System 2 intervention may have some effect when processed first by System 1, but additional System 2 processing will significantly increase its effect.

[Insert Figure 2]

System 1 Intervention. We define a System 1 intervention as one that is intended to be immediately recognized by System 1 with no effortful cognition or deliberate attention paid to the stimuli. System 1 cognition is driven by heuristics formed by long experience (Evans 2014; Kahneman 2011). The heuristics almost instantly match a stimulus to a set of perceptions and behaviors linked to that stimulus. Thus, an intervention striving to influence through System 1 cognition needs a stimulus whose meaning is *instantly* and *intuitively* obvious to the target population without deliberate thought. In the case of computer warning messages, a System 1 intervention can easily be implemented through the icon that accompanies the text.

In this case, our goal is to either instantly stop the user from believing and interacting with the article, or failing that, to provoke cognitive dissonance in the face of confirmation bias. We want to stop a social media user's System 1 from automatically accepting the news article that aligns with their prior beliefs. To trigger a System 1 reaction, we must present the user with a stimulus that invokes the desired intuitive response to stop him or her from believing and automatically accepting the article.

What stimulus invokes an immediate stop reaction? For our target population, adults in the U.S., the answer is obvious: a stop sign. Most U.S. adults drive, and any driver in the U.S. instantly recognizes a stop sign as a signal to stop. Even adults who do not drive often walk around cities or use ride-share applications, causing them to strongly associate stop signs with stopping. Put more formally, most U.S. adults have a System 1 heuristic that interprets a stop sign as a need to stop. Accordingly, when a stop sign appears in the visual field, the immediate System 1 response is "stop."

This System 1 response will shape the perceptions and actions that follow (Kahneman 2011). This immediate stop reaction—a fundamental avoidance response—will influence users' perceptions of the article and the decision to believe it (and subsequent actions such as reading, liking, commenting, and sharing). We theorize that the stop sign will make users want to stop their current action—unquestioningly accepting the article—and thus they will be less likely to believe it. Therefore:

H2a: *Fake news interventions designed to produce a negative System 1 reaction will negatively influence the believability of news articles on social media.*

System 2 Intervention. We define a System 2 intervention as one that is designed to present information that, once processed by the effortful cognition of System 2, is likely to influence perceptions or actions. It is an intervention whose meaning needs to be studied and integrated with the information in the recipient's existing mental model before it has its full ability to influence perceptions and/or actions.

Processing text arguments usually requires System 2 cognition because understanding text and connecting those arguments to prior knowledge requires deliberate attention to detail (Evans and Stanovich 2013; Kahneman 2003; Kahneman 2011). It is simple to read a piece of text and pay little attention to its meaning by relying on System 1. However, for detailed information to be understood and integrated into the user's mental model, System 2 cognition is required (Evans and Stanovich 2013; Kahneman 2003; Kahneman 2011). Thus, we use the text portion of the message to implement our System 2 intervention.

Such a System 2 intervention is a persuasive argument that has the strongest effect when the user elaborates on the message (Petty and Cacioppo 1986); i.e., invokes their System 2 cognition to critically read the message and engage in sufficient cognition to understand its meaning in context (Kahneman 2011). The intervention has to create sufficient cognitive dissonance to motivate the user to resolve the inconsistency between an article that confirmation bias induces the user to believe and the flag that tells the user not to. The intervention needs to be strong enough to pull them out of their state of effortless processing associated with the hedonic mindset.

The Facebook's flag used the text "Disputed by 3rd Party Fact-Checkers," which had little effect on users (Moravec et al. 2019; Pennycook and Rand 2017). We theorize that a fake news flag that uses a more direct language can create stronger cognitive dissonance. The purpose of the flag is to mark articles that third-party fact checkers have determined to be false and deceptive. The term "fake news" has become more common in the past few years, and many social media users are wary of fake news articles. Therefore, we theorize that for many in our target population (U.S. adults), attaching a fake news flag that uses the word "fake" to a news article is likely to be more persuasive when fully considered by System 2 cognition.

We theorize that System 2 cognition will be more skeptical of the flagged news article than System 1 cognition because System 2 is influenced less by confirmation bias than System 1 (Kahneman 2011). This

will cause a decrease in the believability of the article. If the text presented by the flag is persuasive enough, then System 2 will resolve this cognitive dissonance by concluding that the flag is true and the article is false. On the other hand, if the text in the flag is not sufficiently persuasive, System 2 will resolve the dissonance by concluding that the fake article is true and that the flag is wrong. Therefore:

H2b: *Fake news interventions designed to produce a negative System 2 reaction will negatively influence the believability of news articles on social media.*

Combined Intervention. We argued above that our System 1 and System 2 inventions are designed to act *primarily* through two different theoretical routes. Although both System 1 and System 2 cognition can be used, they are separate and distinct. Therefore, an intervention that is designed to influence both will be more effective than an intervention designed to influence only one. In other words, our System 1 and System 2 interventions are likely *complements,* not *substitutes*. Thus, the combination of both interventions will result in the strongest negative effect on the believability of articles:

H2c: *Fake news interventions designed to produce a negative System 1 reaction and a negative System 2 reaction will more negatively influence the believability of news articles on social media compared to either intervention alone.*

If the two interventions do not act through separate theoretical routes, then a combined intervention is likely to have little effect over the separate interventions. Thus, empirical data that support H2c would also suggest that the two interventions indeed act through two different theoretical routes.

Effects on Behavior

Thus far, we have focused on how social media users assess the believability of articles. There are also many ways users can engage with articles. Believability affects whether users choose to read an article, like it (by clicking on *Like* button), share it, or comment on it; users are more likely to read, like, share or comment on articles they believe to be true (Kim and Dennis 2019).

Preexisting opinions also influence these behaviors (Kim and Dennis 2019). A user is more likely to read an article that is congruent with his or her prior opinions due to confirmation bias (Ask and Granhag 2005; Moravec et al. 2019). Social media users are seeking entertainment (Johnson and Kaye 2015;

Sledgianowski and Kulviwat 2009), and viewing information that supports one's opinions is more enjoyable than viewing information that challenges them. Thus, users will be more likely to read articles that support their preexisting opinions. Each of these actions is separate and distinct; many people like or share an article without reading it (Gabielkov et al. 2016), though few users engage with an article beyond reading (Hampton et al. 2012; Lee et al. 2016), perhaps because liking, sharing and commenting require more commitment to the article than reading (Kim and Yang 2017; Muntinga et al. 2011). The predominant behavior of passively consuming content can be attributed to users' concerns related to privacy, social reputation and risk (Acquisti and Gross 2006; Eisingerich et al. 2015; Kietzmann et al. 2011).

The choice to engage with an article can be influenced by an emotional reaction (i.e., System 1 cognition) or the consideration of the information it contains (i.e., System 2 cognition) (Kim and Yang 2017). Liking is driven more by emotion (System 1) and commenting more by consideration of information (System 2) (Kim and Yang 2017). Liking is closely associated with System 1 cognition as liking is easier than commenting and sharing (Sumner et al. 2018). The act of clicking *Like* typically indicates positive sentiment and does not involve critical thinking; rather, it is an emotional and gut-level decision (Kim and Yang 2017; Sumner et al. 2018). Sharing is influenced by both emotion and more deliberate consideration of information (Kim and Yang 2017). Sharing is not only linked to greater commitment, but it is also linked to a user's self-presentation (Kaur et al. 2019; Rui and Stefanone 2013; Van Dijck 2013) and leaves users open to public judgment (Kaur et al. 2019). When public judgment is possible, users will be more cautious of what they share than of what they like. Therefore, users will want to more critically consider what they share than what they like, and this critical consideration comes during System 2 processing.

Believability and confirmation bias are the primary factors that influence whether users read, like, share, and comment on articles (Kim and Dennis 2019). Nonetheless, because these behaviors are influenced in different ways by System 1 and System 2 cognition, System 1 and System 2 interventions designed to influence believability may have effects over and above their effects on believability. A System 1 intervention will likely have a strong effect on liking, which is commonly driven by System 1. Similarly, a System 2 intervention will have a direct effect on commenting which is more strongly driven by System

2. A combined intervention would have a significant effect on sharing which is driven by both System 1 and System 2; the automatic and emotional reaction (System 1) followed by critical consideration using System 2 processing would affect the likelihood of sharing. Thus:

H3: *Over and above the effects of believability, (a) a System 1 intervention will have a direct effect on liking; (b) a System 2 intervention will have a direct effect on commenting; and (c) a combined intervention will have a direct effect on sharing.*

PRELIMINARY TEST

We first need to test that our interventions worked as primarily intended: The stop sign icon (i.e., our System 1 intervention) should be instantaneously understood by System 1, with additional processing by System 2 having little effect. By contrast, the text (i.e., our System 2 intervention) may have some effect when processed by System 1, but subsequent processing by System 2 should significantly increase its effect. Our preliminary test was designed to verify this relationship.

Method

We recruited 44 adult participants in the United States from Mechanical Turk. Researchers in information systems and marketing have found online crowdsourcing markets (OCMs) to be as good or better than student samples at approximating the U.S. population (Steelman et al. 2014). We filtered participants to those that had completed at least 500 tasks with an approval rate of 95% or above, as suggested (Peer et al. 2014; Steelman et al. 2014). About 57% of our participants were female, and 64% were Caucasian. The participants were diverse in age: 9% did not disclose age, 25% were 24−34, 45% were 35−44, 5% were 45−54, 14% were 55−64, and 2% were 65 and older. About 9% of participants did not disclose education information, 14% of participants never attended college, 20% attended some college, 18% completed a two-year degree, 25% completed a four-year degree program, and 14% had a graduate degree. More than 63% used Facebook every day; 27% used it once a week or less. About 16% were Democrats, 66% Republicans, and 18% independent.

Each subject saw both our System 1 intervention (the stop sign) and our System 2 intervention (the "Declared Fake" text). The interventions were displayed for only 1 second (sufficient for System 1

cognition, but too short for System 2 cognition) or 5 seconds (sufficient for System 2 cognition). The time and invention order were randomly assigned; see Figure 3. In the introduction, the subjects learned what we mean by label (i.e., flag), icon and text. To separate the effect of our two interventions, the text part of the label was blurred out when the icon was visible, and vice versa; see Figure 4.

[Insert Figures 3 and 4]

We measured the effectiveness of the label (using a 5-point scale: *How effective is the icon/text (as part of the label) at indicating fake news?* 1= not at all, 5= extremely). For the 1-second treatment group, the subjects were instructed that the label would disappear quickly just to give them an impression of it and were asked to answer the question with their instinctive gut reaction as quickly as possible. For the 5-second treatment group, subjects were instructed that the label would appear for five seconds and were not asked to answer based on their gut reaction. During the debrief survey, we asked the subjects about their overall belief in labels to control for idiosyncrasies in how they perceived warning labels in general (using a 5-point scale: *How effective is any such label at indicating fake news?* 1= not at all, 5= extremely). Also, to verify that the subjects had seen the label, we asked them to describe the icon and the text. Across both 1-second and 5-seconds treatment groups, most subjects were able to correctly recall what they saw.

Results

The mean values in Table 1 suggest that the System 2 intervention was more effective when given more time to process it whereas that was not the case for the System 1 intervention. Table 1 also shows the statistical results using multilevel mixed-effects linear regression in Stata with the 1-second System 1 icon as the baseline. Compared to the baseline, greater time to process the System 1 icon did not yield any statistically significant effect. A Wald test of the System 2 text coefficients shows a statistically significant difference between the System 2 intervention for 1 second and 5 seconds ($p = 0.024$). Thus, we conclude that the interventions work as hypothesized: the icon designed to have its primary effect when processed by System 1 does have its primary effect when processed by System 1 (and System 2 processing adds no value), and the text designed to have its primary effect when processed by System 2 does have this effect (because System 2 processing adds significant value over and above System 1 processing).

[Insert Table 1]

MAIN STUDY

Method

Participants. For our main study, we recruited 398 participants from a Qualtrics panel of adults in the United States. About half of our participants were female, and 81% were Caucasian. The participants were diverse in age: 8% were younger than 24, 20% were 25−34, 17% were 35−44, 18% were 45−54, 23% were 55−64, and 15% were 65 and older. About 30% of participants never attended college, 28% attended some college, 11% completed a two-year degree, 23% completed a four-year degree program, and nearly 8% had a graduate degree; this education pattern is similar to the U.S. population as a whole (Ryan and Bauman 2016). More than 74% used Facebook everyday (14% used it once a week or less). About 43% were Democrats, 39% Republicans, and 18% independent.

Task. Participants viewed 12 news headlines drawn from Kim and Dennis (2019) and reported their perceptions; see Table 2. At the time of our study, all the headlines were still relevant and appropriate for our purpose. The headlines were designed to avoid major differences in the type and magnitude of feelings they would generate. Half of the headlines were designed to appeal to those with left-leaning political ideologies and the other half designed for right-leaning participants. The headlines appeared as they would on Facebook, with a headline, image, poster and source. We used a single gender-neutral name for the poster and 12 plausible, fabricated news sources that we verified as nonexistent (*HotNews.com*, *NewsDesk.com, NewsMedia.com*, *NewsStand.com, NewsToday.com, NewsUnion.com*, *SocialNews.com*, *NewsHeadlines.com*, *TheNationalNews.com, TheNewsRoom.com, TodaysNews.com*, and *TopNews.com*). The headlines were designed to be similar but were randomized in presentation order and treatments to control for any headline-specific effects (e.g., text, news sources, and images).

[Insert Table 2]

Treatments. This experiment included both within-subject and between-subjects treatments; see Figure 5. A within-subject research design is useful to control differences among participants (Gueorguieva and Krystal 2004). The awareness training treatment was within-subject with all subjects participating first

without training and then after training; the order of training could not be randomized because, once a subject was trained, he/she could not be "untrained." The type of flag was the between-subjects treatment to avoid any confusion that may arise from mixing several different types of flags. In other words, each participant saw only one type of flag.

[Insert Figure 5]

Participants first viewed the four control treatment headlines without flags of any type (see Figure 6a) intermixed with their first set of four treatment headlines (with flags). They were then trained on the meaning of the flag and viewed the last four headlines (with same the flags).

All subjects participated in the control treatment and were randomly assigned to one of the four flagging treatments: (i) Facebook, (ii) System 1, (iii) System 2, or (iv) Combination of System 1 and System 2. All flags contained both an icon and a text (i.e., explanation) as this is the usual format of a computer warning message (Anderson et al. 2016; Vance et al. 2018). The first treatment used Facebook's flag design (Figure 6b); the icon and text in the original Facebook's flag was used as the benchmark to provide ecological validity. The second treatment used the System 1 intervention, which replaced the caution sign used by Facebook with a stop sign (Figure 6c). The third treatment was the System 2 intervention, which changed the text from "Disputed by 3rd Party Fact-Checkers" to "Declared Fake by 3rd Party Fact-Checkers" (Figure 6d). The final treatment had both System 1 and 2 interventions (Figure 6e).

[Insert Figure 6]

The awareness training was implemented as a separate page that explained the purpose of the flag, which was to notify users that the article they were viewing was determined to be fake by third-party fact-checkers, such as Politifact or Snopes (Schaedel 2017). As a manipulation check on the effect of the training, we asked the participants whether they noticed the flag and how many they remembered seeing before the training. After introduction and continued exposure to the flag, we asked their confidence in the use of the flag on social media. Nearly 75% of participants rated the flag as neutral or greater in their confidence in the flag. Here is an example of the awareness training message for System 2 intervention:

From the previous 8 articles, did you notice that some of them had the "Fake" label at the bottom?

Facebook is working with 3rd party fact-checkers (e.g., ABC News, Politifact, FactCheck, Snopes, and the Associated Press) to assess the reliability of different news articles and sources. Users may report a story as fake, or Facebook's internal monitoring system may detect suspicious articles. Such articles would then be verified by fact-checking organizations, and at least two of them have to agree before the label is applied.

Control Variables. Confirmation bias has been found to influence users' beliefs in social media articles (Kim and Dennis 2019), so we included it as a control variable. We used the same measure as Kim and Dennis (2018): a combination of two self-reported items for each headline. The first was the participant's perceived importance of the headline (using a 7-point scale: *Do you find the issue described in the article important?* 1= not at all, 7= extremely). The second was the participant's position on the headline (−3= extremely negative to +3= extremely positive). Confirmation bias was measured by multiplying the two variables together, creating a scale from −21 to +21. Thus, we capture both the *direction* (agree/disagree) and the *magnitude* (strongly/weakly) in the fit between one's preexisting beliefs and the articles. If the absolute value of either variable is low, the effect of confirmation bias will be small. Only when both are high will confirmation bias have a strong effect. In addition to confirmation bias, we also included demographic variables (e.g., age, gender, education, etc.).

Dependent Variables. There are two categories of dependent variables. The first is the primary dependent variable: the believability of the headline. This was an average of three items on seven-point Likert scales adapted from (Beltramini 1988): *How believable do you find this article, How truthful do you find this article, How credible do you find this article*. Cronbach's alpha was adequate at 0.95.

The second is the extent of user engagement with the headline. Specifically, we measured what actions the participant would take and how likely the participant would be to: *Read*, *Like*, *Post a supporting comment*, *Post an opposing comment* and *Share*. Each action was assessed separately.

Results

Believability. Table 3 presents the treatment level means and standard deviations for believability. We used multilevel mixed-effects linear regression with random intercepts in Stata to analyze the effects of the treatments on believability. The baseline was the control treatment, and the results are reported in Table 4.

[Insert Tables 3 and 4]

The results show that all flags have significant negative effects on believability (see Table 4). H1 argued that training on the meaning of the flag will increase the flag's effectiveness in nudging users to believe fake news articles less on Facebook. Both the mean believability in Table 3 and the coefficients in Table 4 support this hypothesis. We also used a mixed level model after aggregating the data across all four flagging treatments and found a significant effect of training on how the flags influenced believability ($\beta = -0.283, p < 0.001$). H1 is supported.

We tested the effects of training on each flag separately and report the pairwise effects of all four flags; see Table 5. Training had a significant effect for Facebook's Flag ($p = 0.002$), the System 2 intervention ($p = 0.007$), and the combination intervention ($p < 0.001$), but not for the System 1 intervention. This pattern of results provides evidence of instantiation validity (Lukyanenko and Parsons 2015). Training is likely to have little effect on a flag that is designed to primarily act through System 1 cognition because System 1 relies on heuristics that are not affected by training (Dennis and Minas 2018). System 1 does not consider what we have been taught nor the advice received from others, but rather uses stimulus-response pairs created from past experiences (Dennis and Minas 2018). Training should not influence this unless the training is based on a repeated realistic experience that creates a new stimulus-response pair (Dennis and Minas 2018; Kahneman 2011). However, training should affect System 2 cognition because System 2 cognition can take time to consider knowledge stored in memory that is shaped by training and integrate it with other factors. Our results show that training affected the flagging treatments theorized to operate primarily via System 2 cognition but did not affect the treatment theorized to work mainly through System 1.

[Insert Table 5]

It is also interesting that Facebook's flag benefitted from training whereas our System 1 intervention did not. Facebook's flag has elements that act through both System 1 and System 2 processing (i.e., its exclamation mark and "disputed" text). They are just weaker than our interventions. For Facebook's flag, the effect of training—the increase in the effect that comes from the System 2 element—is significant

when compared to the overall effect of the flag. However, for our System 1 intervention, the effect that comes from the System 1 element is strong, making the effect of training less obvious.

H2 argued that flags designed using a) a System 1 intervention and b) a System 2 intervention would have a negative effect on believability. The results in Table 4 show that H2a and H2b are supported. H2c argued that the combined effects of a System 1 intervention plus a System 2 intervention would be stronger than either alone. We conducted a series of Wald tests to compare the flags; see Table 6. Before training, the combination intervention had a stronger effect on believability than the System 1 intervention ($p < 0.001$) and the System 2 intervention ($p = 0.008$). The same was true after training: System 1 intervention ($p < 0.001$) and the System 2 intervention ($p < 0.001$). Thus, H2c is supported. The most influential flag was the combination of System 1 and System 2 interventions, having a large effect size and a stronger effect on believability than all other flags, both before and after training. This support for H2c further corroborates that the two interventions operate through different cognitive routes, as we theorized.

[Insert Table 6]

User Actions. H3 examined the effects on users' actions; that is, the choice to read, like, post a supporting comment, post an opposing comment, and share; see Table 7. As in past research (Kim and Dennis 2019), we found that believability and confirmation bias consistently affected all five actions.

[Insert Table 7]

H3a hypothesized that a System 1 intervention would affect liking, over and above its effects through believability. There was a significant effect before training but not after, so H3a is partially supported. H3b hypothesized that a System 2 intervention would influence commenting, over and above believability. There was a significant effect on posting a supporting comment after training, but no other effects, so H3b is partially supported. H3c hypothesized that a combined intervention would influence sharing, over and above believability. There was a significant effect on sharing after training, but not before; H3c is partially supported. We also note that, after training, the combined intervention consistently reduced the likelihood that users would engage in all five actions, over and above its effects through believability.

POST-HOC TEST

In the main study, the awareness training always appeared in the middle of the experiment, and all post-training headlines appeared with a flagging treatment (Figure 5). This design left a couple of open questions about the effectiveness of the training. Perhaps, time and exposure to the headlines and labels increased skepticism, rather than the training causing the decrease in belief. We note that the headlines we used were previously studied and showed no evidence of order effect (Kim and Dennis 2019); order of headline appearance had no material influence on believability. While this gives us some assurance that we can probably rule out order effect, it is still worth verifying as the experiments are dissimilar. Second, it is possible that the awareness training stimulated the subjects to become more skeptical about all headlines—flagged or not—instead of boosting the effect of flags. If this is the case, then all headlines that follow the training would be perceived less believable. However, because of the design of the main study (i.e., all post-training headlines were flagged), we cannot examine this possibility. To address this concern, we ran a post-hoc study to answer these remaining questions and to precisely tease out the effect of training on the flags. While we focused more on the effects of System 1 and System 2 enhancements to Facebook's flag in the main study, in the post-hoc test, we place our focus squarely on the effect of training.

Method

We recruited 84 adult participants in the U.S. from Mechanical Turk. We selected participants in the same way as in our preliminary test. About 58% of our participants were female, and 76% were Caucasian. The participants were diverse in age: 1% were younger than 24, 29% were 25−34, 24% were 35−44, 25% were 45−54, 15% were 55−64, and 6% were 65 and older. About 12% of participants never attended college, 23% attended some college, 18% completed a two-year degree, 34% completed a four-year degree program, and 13% had a graduate degree; this education pattern is similar to the U.S. population as a whole (Ryan and Bauman 2016). More than 78% used Facebook everyday (5% used it once a week or less). About 56% were Democrats, 23% Republicans, and 21% independent.

For this study, eight headlines were sufficient, so we chose the first eight from the 12 headlines used in the main study; see Table 2. Similar to the main study, the headlines were evenly split in their ideological appeal, and the headlines were randomized in order and treatment to control for any headline-

specific effects. The participants viewed the eight news headlines and reported their perceptions, which we use to construct the same variables (e.g., believability and confirmation bias) used in the main study.

Similar to the main study, this experiment included both within-subject and between-subjects treatments; see Figure 7. The training message was the same as in the main study. Unlike the main study, however, the awareness training was a randomly assigned between-subjects factor to rule out an order effect or overall increased skepticism. The subjects in the training treatment were shown the awareness training message before all headlines whereas those in the control treatment did not receive any training. All subjects saw four flagged headlines and four control headlines in a random order. Our focus in this post-hoc test was on the effect of training, so we used the Facebook's flag (the benchmark) without any of our System 1 and System 2 interventions. The variables are the same as those used in the main study.

[Insert Figure 7]

Results

We used multilevel mixed-effects linear regression in Stata; see Table 8. The table shows that the awareness training's main effect is not significant, suggesting that the training did not induce the subjects to become more skeptical of headlines if they were not flagged. As expected, flagging had a significant main effect in reducing believability. Most notably, the interaction effect of flagging and training significantly reduced believability. This shows that beyond the influence of the flag itself, training users on the meaning of the flag can significantly increase the influence of the flag.

[Insert Table 8]

DISCUSSION

Fake news on social media, and Facebook's attempt to flag articles that fact-checkers determined to be false, have received much attention. In this paper, we investigated the effects of different presentation formats for fact-checking results, as well as the effect of training about the flag's meaning on the extent to which social media users believed articles and were likely to read, like, comment on, and share them.

A short awareness training message made users less likely to believe articles flagged as fake. For three interventions, the training message led to a statistically significant difference in the belief of articles;

the effect sizes ranged from small to large. This shows that brief awareness training about fake news flags provides a noticeable benefit. As theorized, we found that the intervention designed to primarily influence System 1 cognition did not become more influential when users were trained on it. This result follows from the theoretical argument that System 1 cognition is based on experiential patterns that are learned by experience, not given as advice or training from others (Dennis and Minas 2018; Kahneman 2011).

The design of the flagging intervention had a significant impact on the believability of the news article. We proposed and tested two separate types of interventions, one designed to have its primary effect when processed using System 1 cognition and the other when processed with System 2 cognition. Both System 1 and System 2 interventions were effective, and the combination intervention was significantly more influential than both. The System 1 and System 2 interventions had small effect sizes before training, and the System 2 intervention had a medium effect size after training. The System 1 and System 2 combination intervention had a medium effect size before training and a large effect size after training.

Thus, a combination intervention is most effective in influencing perception. Presenting users with a System 1 intuitive intervention and a System 2 persuasive argument intervention had the strongest influence on the headline's believability. Thus, providing an instantly recognizable stop sign combined with more compelling text improved users' ability to responsibly consume and react to news on social media.

The combination intervention had a greater effect *before the users were trained* than either individual intervention *after the users were trained*. Thus, we believe the best approach is to design a more impactful fake news flag that is intended to directly affect both System 1 cognition and System 2 cognition. Brief awareness training increased the effect of the combination intervention by 50% (see Table 5), so we believe that regular awareness training on the such an intervention is also important.

The limitations of this research are those typically observed in experiments. We used artificial news stories, and participants may have been in more of a utilitarian mindset than they typically are in when using social media, and therefore more likely to think critically regardless of the interventions. That would cause us to underestimate the influence of the System 2 intervention compared to actual use, giving us a conservative estimate. Lastly, the effectiveness of the interventions was tested during a relatively short time

period. The long-term effect of fake news warning labels and awareness training remains an open question.

Implications for Research

The first implication for future research is that flagging fake news can be effective. However, to strengthen these results, we need more research on fact-checking and on ways to present the results of fact-checking so that they would have a strong influence on users' beliefs and behavior. Considering the typical limitations of lab experiments, it may be valuable to conduct research in more naturalistic settings that are less likely to trigger a utilitarian mindset and are able to allow users to dynamically interact with the environment. We presented two separate approaches to the design of a fake news flag, and our findings indicate that the two interventions act though two separate theoretical processes. To provide a more specific design prescription, we need more research on each separate intervention as well as their combined effects.

Second, our research shows that these flags designed to work under System 1 or System 2 cognition influenced users' perceptions on social media. Our research used behavioral methods to draw theoretical conclusions about user cognition. However, without neuroscience, we cannot conclusively prove that the flags used in this study altered cognition. As in past behavioral research, we observe the behavioral impacts that our theorized changes in cognition would lead to, but absent direct measures, we can only indirectly infer the level of cognitive efforts exerted by users. A more direct approach using neuroscience methods (e.g., EEG) may reveal the changes in cognition. "There is no one part of the brain that either of the systems would call home," but there are measurable telltale signs that indicate an increase in cognitive effort (e.g., changes in the brain's overall metabolic demands, pupil dilation, attention level) (Kahneman 2011).

Third, we need more research looking at different interventions designed to reduce the belief in fake news on social media. For example, what are the bare essential attributes of flags necessary to induce a strong gut reaction and/or critical thinking? At what point does the effect begin to diminish? The flag design that Facebook tested may appear separate from the news item and could be easily ignored by users. What static and dynamic (i.e., changing over time) design elements can we use to draw more attention?

Fourth, alternative approaches such as rating *sources*—not *articles* as we did in this study—may be a promising solution that can address the issue of delayed aspect of fact checking individual articles; fact

checking usually occurs after the articles are published and circulated (Shao et al. 2016), after the fake articles have already caused harm. Similar to seller ratings on retailer sites, the sources can be continually evaluated based on prior articles, and the resulting source ratings can be associated with any subsequent articles posted by the sources, thereby avoiding the delayed aspect of fact-checking solutions. Source ratings could also serve as indicators of reliability, as those sources that pop up and spread fake news would either have no rating or negative ratings. Analogous to how sellers with no transaction history are viewed less reliable than reputable sellers, lack of ratings could also provide a useful signal (Hsiao 2018; Issa 2017).

Fifth, our results showed that brief awareness training had significant medium-sized effects for both Facebook's flag and our System 2 flag. Our training intervention was minimal, requiring the user to read a brief announcement, which is not unusual on social media platforms. Training users on better use of social media and consumer technologies has often been overlooked, likely because we expect users to know the tools they use regularly, and because traditional classroom training has little place in the life of consumer technology users. We need more research on brief training messages designed to improve consumers' use of social media. Considering how training is not a normal part of social media use, we should be judicious in how we decide on the flow (intrusive versus non-intrusive) and the frequency of the training messages.

Finally, we adapted recent research from psychology into the information systems domain. We found that applying theories of System 1 and System 2 cognition to the design of information system interventions can better enable responsible use of social media. We encourage other researchers to integrate System 1 and System 2 cognition into their research and to look beyond the assumption that users expend deliberate effort when consuming information online. What the fake news epidemic demonstrates is clear; social media users rely on more than facts and evidence when consuming and engaging with articles.

Implications for Practice

This research has practical implications for social media platforms, social media users, and news organizations. First and foremost, our results show that platforms can reduce the extent to which their users fall for and spread fake news articles by deploying a better designed fake news flag. Facebook already had an interface and process in place to flag fake news but withdrew the flag because it was ineffective (Meixler

2017). By using a flag that is designed to provide strong effect when considered by both System 1 and System 2 cognition, Facebook (and other social media platforms) can induce a large effect in helping its users better avoid fake news with a relatively small change in their interface and little change to the process.

Second, we note that the stop sign and "declared fake" text used in our studies are just one instantiation based on theory to test our hypotheses. The best implementation may be different for different cultural and legal contexts. For instance, there may have been legal reasons why Facebook avoided the term "fake," which may have opened them to libel lawsuits, although Facebook has never stated this reasoning.

Third, social media platforms can reduce the spread of fake news by providing a brief awareness training message that is presented at the top of users' news feeds. Periodic training messages are likely to help users better understand the meaning of the fake news flag and other aspects of social media.

Social media users share some obligation to responsibly use the technology and to not mislead others. Unfortunately, several research articles suggest that the users are not as proficient at detecting fake news as they believe, and the environment on Facebook and Twitter does little to help the users discern truth from fiction, causing fake news to spread rapidly (Vosoughi et al. 2018). It is difficult for users to detect fake news because of the nature of social media, so it is important for platforms to do more to help their users better respond to news—both real and fake—on social media.

CONCLUSION

In conclusion, we found that a flag designed to appeal to both *sense* (i.e., System 1's gut-level reaction) and *sensibility* (i.e., System 2's deliberate cognition) was effective in cautioning users against believing and spreading fake news. Brief awareness training also increased the effectiveness of fake news flags, especially for the combination flag. Our work demonstrates how both theoretical cognitive processes can be used—separately or together—in the design of fact-checking results presented to social media users.

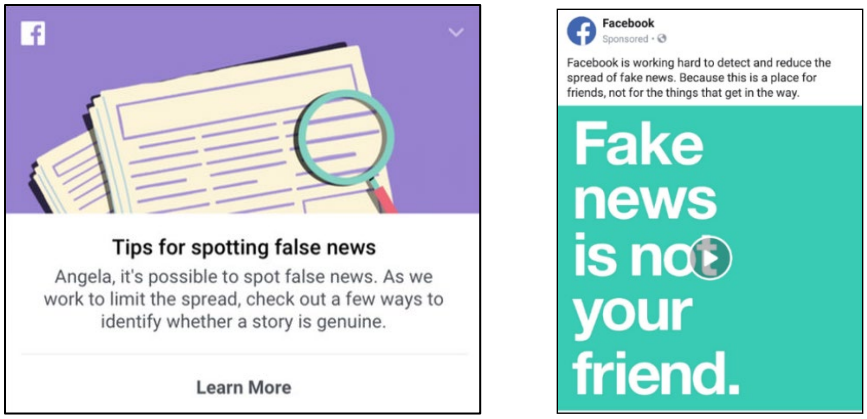

Figure 1: Examples of Awareness Training via Facebook Tips and Announcements

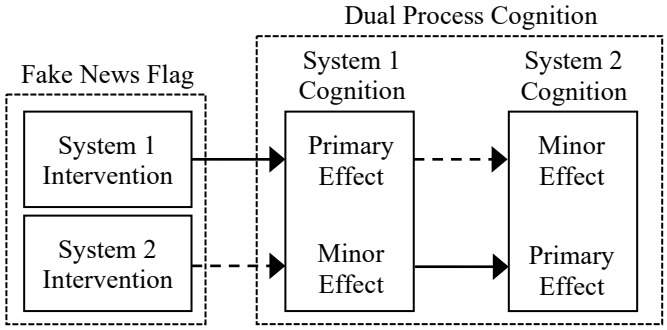

Figure 2: System 1 and System 2 Interventions

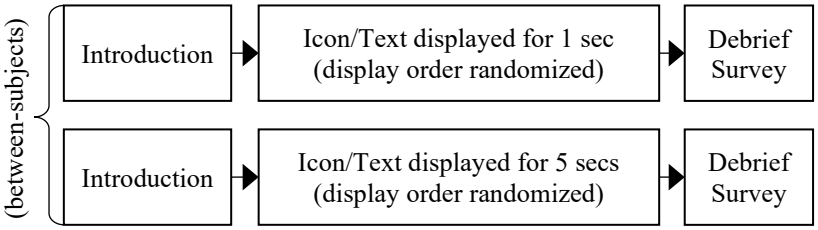

Figure 3: Design of the Preliminary Test Study

(a) System 1 Intervention

(b) System 2 Intervention

Figure 4: Sample of Treatments Used in Preliminary Test

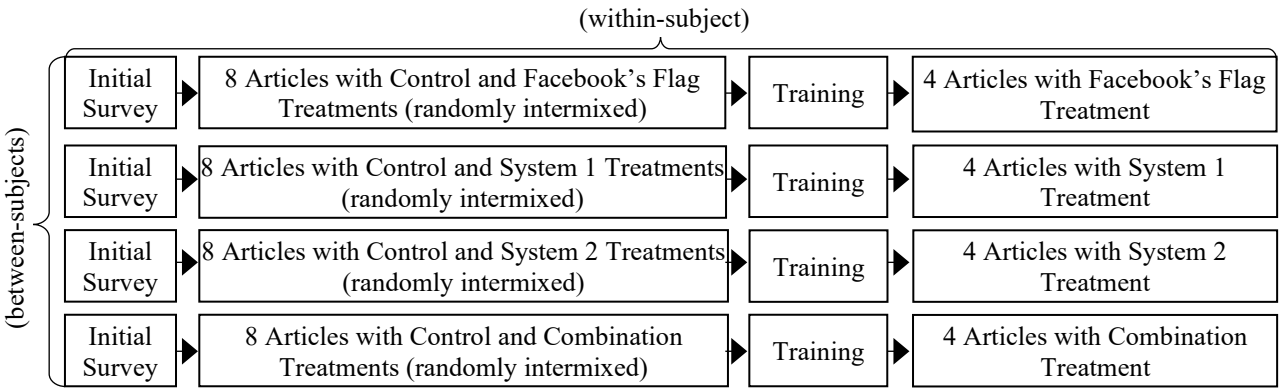

Figure 5: Design of the Main Study Experiment

(a) Control

(b) Facebook's Flag

(c) System 1 Intervention

(d) System 2 Intervention

(e) Combination Intervention

Figure 6: A sample of article (a) in the control format, (b) with the Facebook's flag, (c) with the System 1 intervention, (d) with the System 2 intervention, and (e) with the combination of System 1 and System 2 interventions.

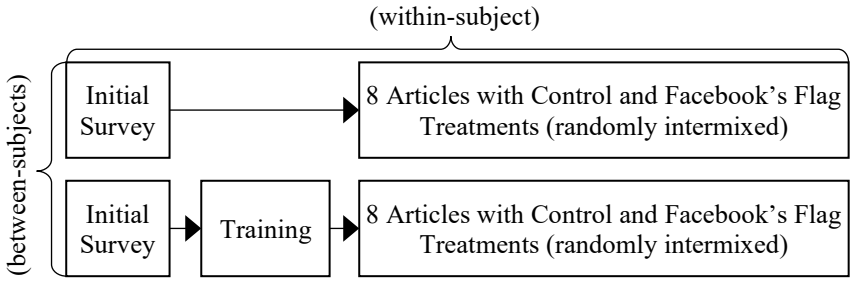

Figure 7: Design of the Post-Hoc Study

Table 1: Statistical Results for Label Effectiveness

Independent Variables		**Mean**	**SD**
System 1 Intervention – 1 second (the baseline)	—	3.318	1.086
System 1 Intervention – 5 seconds	−0.263	2.955	1.253
System 2 Intervention – 1 second	0.136	3.455	1.262
System 2 Intervention – 5 seconds	0.828**	4.045	0.844
Control Variables			
Overall Belief in Labels	0.408***		
Female	−0.082		
Working Age (25-64)	−0.628		
Retirement Age (65+)	−0.054		
Bachelor's Degree	−0.041		
Graduate Degree	0.014		
FB Use > Once A Day	0.342		
FB Use > Once A Week	−0.194		
Democrat	−0.446		
Republican	−0.052		

Note: Estimated coefficients and statistical significance.
***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table 2: The Headlines Used in the Experiments (Based on Kim and Dennis (2019))

- Planned Parenthood Receives a Sum of $1,000,000 Donation from Crowd Sourcing
- Girl Scouts are Planning an Organization-Wide Fundraiser for Planned Parenthood
- Planned Parenthood Visits Campuses to Educate Young Women about the Importance of Having a Choice
- Universities Connect their Healthcare Systems with Planned Parenthood to Provide Better Care to Students
- State Republicans Introduce New Bills to Allow Abortion Only After a Long Monitoring Period
- The State of Nevada Strengthens the Restriction on Abortion and Contraception
- Planned Parenthood Now Required to Provide Classes on Abortion Before Getting Consent for the Procedure
- On-Campus Pro-Life Supporters Significantly Reduce the Number of Abortions among Students
- The Humane Society Foundation Donates $100,000 to Planned Parenthood After Women's March in DC
- A Republican GOP Senator Will Not Vote to Defund Planned Parenthood
- Republicans Pledge to Only Fund National Pregnancy Care Center That Does Not Perform Abortions
- Pro-Life Supporters Rally in Front of Planned Parenthood Nationwide

Table 3: Treatment Level Means and Standard Deviations for Believability

	Before Training		After Training	
Treatment	**Mean**	**Std**	**Mean**	**Std**
Control (No Flag)	4.712	1.811	n/a	n/a
Facebook's Flag	4.446	1.612	4.087	1.585
System 1 Intervention	4.027	1.933	3.983	1.849
System 2 Intervention	4.079	1.813	3.774	1.783
Combination Intervention	3.506	1.802	3.218	1.778

Table 4: Statistical Results for Believability

Awareness training results summarized in Table 5

Flag comparisons summarized in Table 6

Independent Variables		
Without Awareness Training	Facebook's Flag	-0.291^{**}
	System 1 Intervention	-0.499^{***}
	System 2 Intervention	-0.553^{***}
	Combination Intervention	-0.995^{***}
With Awareness Training	Facebook's Flag	-0.621^{***}
	System 1 Intervention	-0.590^{***}
	System 2 Intervention	-0.819^{***}
	Combination Intervention	-1.403^{***}
Control Variables		
Confirmation Bias		0.063^{***}
Female		−0.019
Working Age (25-64)		−0.072
Retirement Age (65+)		-0.434^{*}
Bachelor's Degree		−0.036
Graduate Degree		0.051
FB Use > Once A Day		−0.231
FB Use > Once A Week		0.075
Democrat		0.241
Republican		0.310^{*}

Note: Estimated coefficients and statistical significance.
$^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$.

Table 5: Wald Tests of Within-Subject Interventions (Awareness Training Results)

Within Treatment Comparison Before and After Training	Chi-Square
Facebook's Flag Before (−0.291) versus After (−0.621)	9.47**
System 1 Intervention Before (−0.499) versus After (−0.590)	0.77
System 2 Intervention Before (−0.553) versus After (−0.819)	6.40**
Combination Intervention Before (−0.995) versus After (−1.403)	14.12***

Note: ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table 6: Wald Tests of Between-Subjects Interventions (Flag Comparisons)

Before Awareness Training	Chi-Square
Facebook's Flag (−0.291) versus Combination Intervention (−0.995)	26.13***
Facebook's Flag (−0.291) versus System 1 Intervention (−0.499)	2.48
Facebook's Flag (−0.291) versus System 2 Intervention (−0.553)	3.87*
System 1 Intervention (−0.499) versus System 2 Intervention (−0.553)	0.17
Combination Intervention (−0.995) versus System 1 Intervention (−0.499)	13.42***
Combination Intervention (−0.995) versus System 2 Intervention (−0.553)	10.56**
After Awareness Training	
Facebook's Flag (−0.621) versus Combination Intervention (−1.403)	37.64***
Facebook's Flag (−0.621) versus System 1 Intervention (−0.590)	0.05
Facebook's Flag (−0.621) versus System 2 Intervention (−0.819)	2.21
System 1 Intervention (−0.590) versus System 2 Intervention (−0.819)	3.05
Combination Intervention (−1.403) versus System 1 Intervention (−0.590)	42.10***
Combination Intervention (−1.403) versus System 2 Intervention (−0.819)	21.40***

Note: ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table 7: Results for User Actions

Independent Variables		**Read**	**Like**	**Support**	**Oppose**	**Share**
Believability		0.251***	0.381***	0.212***	0.194***	0.283***
Without Training	Facebook's Flag	0.097	0.059	−0.008	0.099	−0.114
	System 1	−0.157	−0.216*	−0.132	−0.173	−0.243**
	System 2	0.001	0.061	0.052	0.003	0.048
	Combined S1&S2	0.061	−0.130	0.093	0.090	−0.061
With Training	Facebook's Flag	−0.042	−0.076	−0.007	0.000	−0.002
	System 1	−0.218**	0.006	−0.126	−0.018	−0.083
	System 2	−0.138	−0.223*	−0.172*	−0.064	−0.161*
	Combined S1&S2	−0.214*	−0.395***	−0.279***	−0.252**	−0.314***
Control Variables						
Confirmation Bias		0.078***	0.020**	0.064***	−0.042***	0.027***
Female		0.129	0.255	0.138	0.048	0.175
Working Age (25-64)		−0.363	−0.215	−0.146	−0.194	−0.360
Retirement Age (65+)		−0.760**	−0.594	−0.701*	−0.845*	−0.924**
Bachelor's degree		−0.163	−0.032	−0.182*	−0.327	−0.169
Graduate Degree		0.002	0.018	0.046	−0.021	0.088
FB Use > Once A Week		−0.499**	0.569*	−0.441*	−0.491*	−0.575**
FB Use > Once A Day		−0.326	0.573*	−0.290	−0.435	−0.291
Democrat		0.576***	0.666***	0.583***	0.687**	0.578**
Republican		0.302	0.498*	0.388	0.415	0.200

Note: Estimated coefficients and statistical significance. ***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

Table 8: Statistical Results for Believability in the Post-Hoc Study

Independent Variables	
Awareness Training	−0.172
Flagged Headline	−0.932**
Training × Flag	−0.620**
Control Variables	
Confirmation Bias	0.049***
Female	0.037
Working Age (25-64)	−0.670
Retirement Age (65+)	−0.759
Bachelor's Degree	0.181
Graduate Degree	0.058
FB Use > Once A Day	0.509
FB Use > Once A Week	0.506
Democrat	0.372*
Republican	0.264*

Note: Estimated coefficients and statistical significance.
***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$.

REFERENCES

Achtziger, A., and Alós-Ferrer, C. 2013. "Fast or Rational? A Response-Times Study of Bayesian Updating," *Management Science* (60:4), pp. 923-938.

Acquisti, A., and Gross, R. 2006. "Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook," *International workshop on privacy enhancing technologies*: Springer, pp. 36-58.

Allcott, H., and Gentzkow, M. 2017. "Social Media and Fake News in the 2016 Election," National Bureau of Economic Research.

Anderson, B. B., Vance, A., Kirwan, C. B., Eargle, D., and Jenkins, J. L. 2016. "How Users Perceive and Respond to Security Messages: A Neurois Research Agenda and Empirical Study," *European Journal of Information Systems* (25:4), pp. 364-390.

Anderson, C. L., and Agarwal, R. 2010. "Practicing Safe Computing: A Multimethod Empirical Examination of Home Computer User Security Behavioral Intentions," *MIS Quarterly* (34:3), pp. 613-643.

Aronson, E. 1969. "The Theory of Cognitive Dissonance: A Current Perspective," *Advances in experimental social psychology* (4), pp. 1-34.

Ask, K., and Granhag, P. A. 2005. "Motivational Sources of Confirmation Bias in Criminal Investigations: The Need for Cognitive Closure," *Journal of Investigative Psychology and Offender Profiling* (2:1), pp. 43-63.

Auer, M. M., and Griffiths, M. D. 2015. "Testing Normative and Self-Appraisal Feedback in an Online Slot-Machine Pop-up in a Real-World Setting," *Frontiers in psychology* (6), p. 339.

Bago, B., and De Neys, W. 2017. "Fast Logic?: Examining the Time Course Assumption of Dual Process Theory," *Cognition* (158), pp. 90-109.

Bansal-Travers, M., Hammond, D., Smith, P., and Cummings, K. M. 2011. "The Impact of Cigarette Pack Design, Descriptors, and Warning Labels on Risk Perception in the Us," *American journal of preventive medicine* (40:6), pp. 674-682.

Barthel, M., Mitchell, A., and Holcomb, J. 2016. "Many Americans Believe Fake News Is Sowing

Confusion," *Pew Research Center* (15).
Beltramini, R. F. 1988. "Perceived Believability of Warning Label Information Presented in Cigarette Advertising," *Journal of Advertising* (17:2), pp. 26-32.
Bernstam, E. V., Shelton, D. M., Walji, M., and Meric-Bernstam, F. 2005. "Instruments to Assess the Quality of Health Information on the World Wide Web: What Can Our Patients Actually Use?," *International Journal of Medical Informatics* (74:1), pp. 13-19.
Bravo-Lillo, C., Cranor, L. F., Downs, J., and Komanduri, S. 2010. "Bridging the Gap in Computer Security Warnings: A Mental Model Approach," *IEEE Security & Privacy* (9:2), pp. 18-26.
Cerf, V. G. 2016. "Information and Misinformation on the Internet," *Commun. ACM* (60:1), pp. 9-9.
Chauhan, K., and Pillai, A. 2013. "Role of Content Strategy in Social Media Brand Communities: A Case of Higher Education Institutes in India," *Journal of Product & Brand Management* (22:1), pp. 40-51.
Cheung, C. M.-Y., Sia, C.-L., and Kuan, K. K. 2012. "Is This Review Believable? A Study of Factors Affecting the Credibility of Online Consumer Reviews from an Elm Perspective," *Journal of the Association for Information Systems* (13:8), p. 618.
Cotte, J., Chowdhury, T. G., Ratneshwar, S., and Ricci, L. M. 2006. "Pleasure or Utility? Time Planning Style and Web Usage Behaviors," *Journal of interactive marketing* (20:1), pp. 45-57.
Cranor, L. F. 2008. "A Framework for Reasoning About the Human in the Loop," in: *Proc. 1st Conf. Usability, Psychology, and Security (UPSEC 08)*. USENIX.
D'Arcy, J., Hovav, A., and Galletta, D. 2009. "User Awareness of Security Countermeasures and Its Impact on Information Systems Misuse: A Deterrence Approach," *Information Systems Research* (20:1), pp. 79-98.
de Castro Bellini-Leite, S. 2013. "The Embodied Embedded Character of System 1 Processing," *Mens sana monographs* (11:1), p. 239.
de Guinea, A. O., and Markus, M. L. 2009. "Why Break the Habit of a Lifetime? Rethinking the Roles of Intention, Habit, and Emotion in Continuing Information Technology Use," *MIS Quarterly* (33:3), pp. 433-444.
Dennis, A. R., and Minas, R. K. 2018. "Security on Autopilot: Why Current Security Theories Hijack Our Thinking and Lead Us Astray," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* (49:1), pp. 15-38.
Devine, P. G., Hirt, E. R., and Gehrke, E. M. 1990. "Diagnostic and Confirmation Strategies in Trait Hypothesis Testing," *Journal of Personality and Social Psychology* (58:6), p. 952.
Eisingerich, A. B., Chun, H. H., Liu, Y., Jia, H., and Bell, S. J. 2015. "Why Recommend a Brand Face-to-Face but Not on Facebook? How Word-of-Mouth on Online Social Sites Differs from Traditional Word-of-Mouth," *Journal of Consumer Psychology* (25:1), pp. 120-128.
Evans, J. S. B., and Stanovich, K. E. 2013. "Dual-Process Theories of Higher Cognition: Advancing the Debate," *Perspectives on psychological science* (8:3), pp. 223-241.
Evans, J. S. B. T. 2014. "Two Minds Rationality," *Thinking & Reasoning* (20:2), pp. 129-146.
Eysenbach, G., Yihune, G., Lampe, K., Cross, P., and Brickley, D. 2000. "Medcertain: Quality Management, Certification and Rating of Health Information on the Net," *Proceedings of the AMIA Symposium*, pp. 230-234.
Facebook Help Center. 2017. "Facebook Help Center." from https://www.facebook.com/help/
Festinger, L. 1962. *A Theory of Cognitive Dissonance*. Stanford university press.
Gabielkov, M., Ramachandran, A., Chaintreau, A., and Legout, A. 2016. "Social Clicks: What and Who Gets Read on Twitter?," *ACM SIGMETRICS / IFIP Performance 2016*, Antibes Juan-les-Pins, France.
Gottfried, J., and Shearer, E. 2016. *News Use across Social Medial Platforms 2016*. Pew Research Center.
Graves, L. 2016. "Boundaries Not Drawn: Mapping the Institutional Roots of the Global Fact-

Checking Movement," *Journalism Studies*), pp. 1-19.
Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. 2008. "Cognitive Load Selectively Interferes with Utilitarian Moral Judgment," *Cognition* (107:3), pp. 1144-1154.
Gueorguieva, R., and Krystal, J. H. 2004. "Move over Anova: Progress in Analyzing Repeated-Measures Data Andits Reflection in Papers Published in the Archives of General Psychiatry," *Archives of general psychiatry* (61:3), pp. 310-317.
Haile, T. 2014. "What You Think You Know About the Web Is Wrong." *Time*, from http://time.com/12933/what-you-think-you-know-about-the-web-is-wrong/
Hampton, K. N., Goulet, L. S., Marlow, C., and Rainie, L. 2012. "Why Most Facebook Users Get More Than They Give," *Pew Internet & American Life Project* (3), pp. 1-40.
Harsanyi, J. C. 1977. "Morality and the Theory of Rational Behavior," *Social Research* (44:4), p. 24.
Head, A. J., Wihbey, J., Metaxas, P. T., MacMillan, M., and Cohen, D. 2018. "How Students Engage with News," Knight Foundation, Project Information Literacy, pp. 1-53.
Hirschman, E. C., and Holbrook, M. B. 1982. "Hedonic Consumption: Emerging Concepts, Methods and Propositions," *Journal of Marketing* (46:3), pp. 92-101.
Ho, Y.-C., Wu, J., and Tan, Y. 2017. "Disconfirmation Effect on Online Rating Behavior: A Structural Model," *Information Systems Research* (28:3), pp. 626-642.
Housholder, E. E., and LaMarre, H. L. 2014. "Facebook Politics: Toward a Process Model for Achieving Political Source Credibility through Social Media," *Journal of Information Technology & Politics* (11:4), pp. 368-382.
Hsiao, A. 2018. "Ebay Feedback Evaluation Quick Guide." from https://www.thebalancesmb.com/ebay-feedback-evaluation-quick-guide-1139908
Isaac, M. 2016. "Facebook Mounts Effort to Limit Tide of Fake News." *The New York Times*, from https://www.nytimes.com/2016/12/15/technology/facebook-fake-news.html
Issa, E. E. 2017. "Bad Credit Vs. No Credit: Which Is Worse When Trying to Rent an Apartment?", from https://www.huffpost.com/entry/bad-credit-vs-no-credit-w_b_11334012
Johns, G. 2006. "The Essential Impact of Context on Organizational Behavior," *Academy of management review* (31:2), pp. 386-408.
Johns, G. 2017. "Reflections on the 2016 Decade Award: Incorporating Context in Organizational Research," *Academy of Management Review* (42:4), pp. 577-595.
Johnson, T. J., and Kaye, B. K. 2015. "Reasons to Believe: Influence of Credibility on Motivations for Using Social Networks," *Computers in Human Behavior* (50), pp. 544-555.
Johnston, A. C., Warkentin, M., and Siponen, M. T. 2015. "An Enhanced Fear Appeal Rhetorical Framework: Leveraging Threats to the Human Asset through Sanctioning Rhetoric," *Mis Quarterly* (39:1), pp. 113-134.
Kahneman, D. 2003. "Maps of Bounded Rationality: Psychology for Behavioral Economics," *The American Economic Review* (93:5), pp. 1449-1475.
Kahneman, D. 2011. *Thinking, Fast and Slow*. Macmillan.
Kaur, W., Balakrishnan, V., Rana, O., and Sinniah, A. 2019. "Liking, Sharing, Commenting and Reacting on Facebook: User Behaviors' Impact on Sentiment Intensity," *Telematics and Informatics* (39), pp. 25-36.
Kelley, C. A., Gadis, W. C., and Reingen, P. H. 1989. "The Use of Vivid Stimuli to Enhance Comprehension of the Content of Product Warning Messages," *The Journal of Consumer Affairs* (23:2), pp. 243-266.
Khatri, V., Samuel, B. M., and Dennis, A. R. 2018. "System 1 and System 2 Cognition in the Decision to Adopt and Use a New Technology," *Information & Management* (55:6), p. 16.
Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. 2011. "Social Media? Get Serious! Understanding the Functional Building Blocks of Social Media," *Business horizons* (54:3), pp. 241-251.

Kim, A., and Dennis, A. R. 2019. "Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media," *MIS Quarterly* (43:3), pp. 1025-1039.
Kim, C., and Yang, S.-U. 2017. "Like, Comment, and Share on Facebook: How Each Behavior Differs from the Other," *Public Relations Review* (43:2), pp. 441-449.
Kirby, E. J. 2016. "The City Getting Rich from Fake News." *BBC News* 5 December 2016. from http://www.bbc.com/news/magazine-38168281
Knobloch-Westerwick, S., and Lavis, S. M. 2017. "Selecting Serious or Satirical, Supporting or Stirring News? Selective Exposure to Partisan Versus Mockery News Online Videos," *Journal of Communication* (67:1), pp. 54-81.
Koriat, A., Lichtenstein, S., and Fischhoff, B. 1980. "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory* (6:2), pp. 107-118.
Lappas, T., Sabnis, G., and Valkanas, G. 2016. "The Impact of Fake Reviews on Online Visibility: A Vulnerability Assessment of the Hotel Industry," *Information Systems Research* (27:4), pp. 940-961.
Lee, S.-Y., Hansen, S. S., and Lee, J. K. 2016. "What Makes Us Click "Like" on Facebook? Examining Psychological, Technological, and Motivational Factors on Virtual Endorsement," *Computer Communications* (73), pp. 332-341.
Levin, S. 2017. "Mark Zuckerberg: I Regret Ridiculing Fears over Facebook's Effect on Election." *The Guardian*, 2017, from https://www.theguardian.com/technology/2017/sep/27/mark-zuckerberg-facebook-2016-election-fake-news
Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.
Loewenstein, G., O'Donoghue, T., and Bhatia, S. 2015. "Modeling the Interplay between Affect and Deliberation," *Decision* (2:2), pp. 55-81.
Lowrey, W. 2017. "The Emergence and Development of News Fact-Checking Sites: Institutional Logics and Population Ecology," *Journalism Studies* (18:3), pp. 376-394.
Lukyanenko, R., and Parsons, J. 2015. "Information Quality Research Challenge: Adapting Information Quality Principles to User-Generated Content," *Journal of Data and Information Quality (JDIQ)* (6:1), p. 3.
Lyons, T. 2017. "Replacing Disputed Flags with Related Articles." *Facebook Newsroom*, from https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/
Meixler, E. 2017. "Facebook Is Dropping Its Fake News Red Flag Warning after Finding It Had the Opposite Effect." *Time*, from http://time.com/5077002/facebook-fake-news-articles/
Mills, J. 1999. "Improving the 1957 Version of Dissonance Theory," in *Cognitive Dissonance: Progress on a Pivotal Theory in Social Psychology*. Washington, DC, US: American Psychological Association, pp. 25-42.
Mograbi, G. J. C. 2011. "Neural Basis of Decision-Making and Assessment: Issues on Testability and Philosophical Relevance," *Mens sana monographs* (9:1), pp. 251-259.
Mohammed, S. N. 2012. *The (Dis)Information Age: The Persistence of Ignorance*, (1 ed.). Peter Lang Inc.
Moravec, P. L., Minas, R. K., and Dennis, A. R. 2019. "Fake News on Social Media: People Believe What They Want to Believe When It Makes No Sense at All," *MIS Quarterly* (43:4), pp. 1343-1360.
Muntinga, D. G., Moorman, M., and Smit, E. G. 2011. "Introducing Cobras: Exploring Motivations for Brand-Related Social Media Use," *International Journal of advertising* (30:1), pp. 13-46.
Nickerson, R. S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* (2:2), p. 26.
Panger, G. 2018. "People Tend to Wind Down, Not up, When They Browse Social Media," *Proceedings of the ACM on Human-Computer Interaction* (2:CSCW), pp. 133:131-

133:129.
Park, J., Konana, P., Gu, B., Kumar, A., and Raghunathan, R. 2013. "Information Valuation and Confirmation Bias in Virtual Communities: Evidence from Stock Message Boards," *Information Systems Research* (24:4), pp. 1050-1067.
Parkinson, H. J. 2016. "Click and Elect: How Fake News Helped Donald Trump Win a Real Election." *The Guardian*, from https://www.theguardian.com/commentisfree/2016/nov/14/fake-news-donald-trump-election-alt-right-social-media-tech-companies
Peer, E., Vosgerau, J., and Acquisti, A. 2014. "Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk," *Behavior research methods* (46:4), pp. 1023-1031.
Pennycook, G., and Rand, D. G. 2017. "Assessing the Effect of "Disputed" Warnings and Source Salience on Perceptions of Fake News Accuracy," *Social Science Research Network*).
Petty, R. E., and Cacioppo, J. T. 1986. "The Elaboration Likelihood Model of Persuasion," in *Communication and Persuasion*. Springer, pp. 1-24.
Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A., and Menczer, F. 2011. "Truthy: Mapping the Spread of Astroturf in Microblog Streams," *Proceedings of the 20th International Conference Companion on World Wide Web*: ACM, pp. 249-252.
Rui, J. R., and Stefanone, M. A. 2013. "Strategic Image Management Online: Self-Presentation, Self-Esteem and Social Network Perspectives," *Information, Communication & Society* (16:8), pp. 1286-1305.
Ryan, C., and Bauman, K. 2016. "Educational Attainment in the United States: 2015," United States Census Bureau.
Schaedel, S. 2017. "How to Flag Fake News on Facebook." from http://www.factcheck.org/2017/07/flag-fake-news-facebook/
Shane, S. 2017. "The Fake Americans Russia Created to Influence the Election." from https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html
Shao, C., Ciampaglia, G. L., Flammini, A., and Menczer, F. 2016. "Hoaxy: A Platform for Tracking Online Misinformation," Proceedings of the 25th International Conference Companion on World Wide Web, pp. 745-750.
Silverman, C. 2016. "This Analysis Shows How Viral Fake Election News Stories Outperformed Real News on Facebook." *BuzzFeed News*, 2016, from https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook
Silverman, C., and Singer-Vine, J. 2016. "Most Americans Who See Fake News Believe It, New Survey Says." *BuzzFeed News*, from https://www.buzzfeednews.com/article/craigsilverman/fake-news-survey
Sledgianowski, D., and Kulviwat, S. 2009. "Using Social Network Sites: The Effects of Playfulness, Critical Mass and Trust in a Hedonic Context," *Journal of Computer Information Systems* (49:4), pp. 74-83.
Srull, T. K., and Wyer, R. S. 1980. "Category Accessibility and Social Perception: Some Implications for the Study of Person Memory and Interpersonal Judgments." US: American Psychological Association, pp. 841-856.
Stanovich, K. E., and West, R. F. 2000. "Individual Differences in Reasoning: Implications for the Rationality Debate?," *Behavioral and brain sciences* (23:5), pp. 645-665.
Statista. 2018. "Number of Monthly Active Facebook Users Worldwide as of 1st Quarter 2018 (in Millions)."
Steelman, Z. R., Hammer, B. I., and Limayem, M. 2014. "Data Collection in the Digital Age: Innovative Alternatives to Student Samples," *Journal of Consumer Psychology* (23:2), pp. 212-219.
Straub Jr, D. W. 1990. "Effective Is Security: An Empirical Study," *Information Systems Research* (1:3), pp. 255-276.

Sumner, E. M., Ruge-Jones, L., and Alcorn, D. 2018. "A Functional Approach to the Facebook Like Button: An Exploration of Meaning, Interpersonal Functionality, and Potential Alternative Response Buttons," *New Media & Society* (20:4), pp. 1451-1469.
Thatcher, J. B., Wright, R. T., Sun, H., Zagenczyk, T. J., and Klein, R. 2018. "Mindfulness in Information Technology Use: Definitions, Distinctions, and a New Measure," *MIS Quarterly* (42:3), pp. 831-847.
The Wall Street Journal. 2016. "Blue Feed, Red Feed." *The Wall Street Journal*, from http://graphics.wsj.com/blue-feed-red-feed/
Thompson, V. A. 2013. "Why It Matters: The Implications of Autonomous Processes for Dual Process Theories—Commentary on Evans & Stanovich (2013)," *Perspectives on Psychological Science* (8:3), pp. 253-256.
Van Dijck, J. 2013. "'You Have One Identity': Performing the Self on Facebook and Linkedin," *Media, culture & society* (35:2), pp. 199-215.
Vance, A., Jenkins, J. L., and Anderson, B. B. 2018. "Tuning out Security Warnings: A Longitudinal Examination of Habituation through Fmri, Eye Tracking, and Field Experiments," *MIS Quarterly* (42:2), pp. 355-380.
Vosoughi, S., Roy, D., and Aral, S. 2018. "The Spread of True and False News Online," *Science* (359:6380), pp. 1146-1151.
Wakabayashi, D., and Shane, S. 2017. "Twitter, with Accounts Linked to Russia to Face Congress over Role in Election." from https://www.nytimes.com/2017/09/27/technology/twitter-russia-election.html
Wintersieck, A. L. 2017. "Debating the Truth: The Impact of Fact-Checking During Electoral Debates," *American Politics Research* (45:2), pp. 304-331.
Wogalter, M. S., Conzola, V. C., and Smith-Jackson, T. L. 2002. "Research-Based Guidelines for Warning Design and Evaluation," *Applied Ergonomics* (33), pp. 219–230.
Wogalter, M. S., Conzola, V. C., and Vigilante Jr, W. J. 1999. "Applying Usability Engineering Principles to the Design and Testing of Warning Messages," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*: SAGE Publications Sage CA: Los Angeles, CA, pp. 921-925.
Wohl, M. J., Gainsbury, S., Stewart, M. J., and Sztainert, T. 2013. "Facilitating Responsible Gambling: The Relative Effectiveness of Education-Based Animation and Monetary Limit Setting Pop-up Messages among Electronic Gaming Machine Players," *Journal of Gambling Studies* (29:4), pp. 703-717.
Yin, D., Mitra, S., and Zhang, H. 2016. "Research Note—When Do Consumers Value Positive Vs. Negative Reviews? An Empirical Investigation of Confirmation Bias in Online Word of Mouth," *Information Systems Research* (27:1), pp. 131-144.