The Promise and Peril of Real-Time Corrections to Political Misperceptions

R. Kelly Garrett **Brian E. Weeks**
School of Communication
Ohio State University
Columbus, OH 43210, USA
{garrett.258,weeks.311}@osu.edu

ABSTRACT
Computer scientists have responded to the high prevalence of inaccurate political information online by creating systems that identify and flag false claims. Warning users of inaccurate information as it is displayed has obvious appeal, but it also poses risk. Compared to post-exposure corrections, real-time corrections may cause users to be more resistant to factual information. This paper presents an experiment comparing the effects of real-time corrections to corrections that are presented after a short distractor task. Although real-time corrections are modestly more effective than delayed corrections overall, closer inspection reveals that this is only true among individuals predisposed to reject the false claim. In contrast, individuals whose attitudes are supported by the inaccurate information distrust the source more when corrections are presented in real time, yielding beliefs comparable to those never exposed to a correction. We find no evidence of real-time corrections encouraging counterargument. Strategies for reducing these biases are discussed.

Author Keywords
Credibility; Misinformation; Learning

ACM Classification Keywords
H.1.2 User/Machine Systems.

INTRODUCTION

Inaccurate information is notoriously common on the web [49, 50]. Hundreds of false or unsubstantiated claims on a host of topics, from the link between vaccines and autism to the birthplace of the President, can be found in seconds using a search engine or by perusing relevant blogs. Political misperceptions—beliefs about candidates and issues that are not supported by the best available evidence—are particularly prevalent. Survey data indicate that the more people rely on the Internet for political news, especially partisan blogs, the more false rumors they encounter [16]. This is perhaps unsurprising: the Internet is a unique communication medium, characterized by its broad reach, high speed, low cost, and by the persistence of posted information. Taken together, these attributes give voice to individuals and groups who might otherwise lack the necessary resources to share their beliefs publicly, and this exposure can grant false claims an air of legitimacy [23, 49].

There are several mechanisms that animate the flow and acceptance of misinformation [25]. Sometimes inaccurate beliefs result from a misunderstanding or a failure of memory; at other times, they are the product of politically motivated deception. Whatever their source, though, people tend to hold on to inaccurate beliefs, especially when they are consistent with their world view [25] or when they provide an explanation for an otherwise puzzling phenomenon [47]. Even individuals who strive to be impartial and who have no political axe to grind tend to be biased in their response to corrections by virtue of the mental shortcuts that they use when evaluating new information [46]. Fortunately, however, given sufficient evidence, even the most biased audiences can be moved to reject falsehoods [44].

In response to the prevalence of online misinformation, researchers have begun crafting systems designed to help people make sense of the vast array of competing claims. These systems employ a diverse array of techniques, but they share a common goal: to help users distinguish between truth and falsehoods. Of particular interest here are systems such as Dispute Finder [14], which attempt to present corrections to inaccurate information in real time. The appeal of real-time corrections is self-evident: if we can identify or correct a false statement when the user first encounters it, it is at least plausible that we could limit its influence by preventing its acceptance and further dissemination.

This paper concerns the theoretical assumptions on which systems providing real-time corrections are premised. Specifically, it describes an experimental test of the effectiveness of this approach when correcting inaccurate political information. Despite its obvious appeal, psychological theory suggests that real-time correction might not be as effective as it at first appears. There are numerous reasons to think that a system presenting corrections alongside an inaccurate statement may inadvertently provoke users into defending attitude-

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

consistent misperceptions. In other words, there may be conditions under which this strategy could backfire, undermining corrections rather than reinforcing them. Our goal here is to test whether real-time corrections are in fact more effective than corrections that are presented after a delay.

Rather than build a working prototype to test this claim, we construct an experiment that simulates users' experience under different designs. We compare participants who are presented with an inaccurate statement and no correction to those who see a correction after a delay and to those see a message in which disputed information is highlighted and accompanied by a correction. Results indicate that although real-time corrections are sometimes better than delayed corrections, they are *less* effective when the correction poses a threat to individuals' political attitudes. We argue that this raises significant concerns about the approach employed in a number of contemporary designs; however, we also suggest that there are strategies that can help alleviate this problem. We begin by reviewing the current design space.

BACKGROUND AND RELATED DESIGN WORK

Journalists and web developers have responded vigorously to the high profile of falsehoods circulating online. Snopes, one of the earliest web sites to address this issue, provides detailed assessments of many controversial or outrageous claims found online. News organizations have followed suit, offering fact-checking sites that focus more narrowly on newsworthy political misperceptions (e.g., FactCheck.org and PolitiFact). Most recently, some organizations have experimented with crowdsourcing as a means of fact checking the news (e.g., NewsTrust's TruthSquad).

Researchers have taken up the challenge as well, looking for ways to stem the flow of misinformation online by creating technologies that can identify and, in some cases, correct inaccuracies. Not all of these systems concern political misperceptions, but the mechanisms they employ to enhance users' understanding of the information environment are relevant. Computer scientists have developed systems capable of tracking short distinctive phrases, including rumors and misinformation, as they move across the Web, documenting how they evolve and mutate [28, 48]. Other researchers focus specifically on identifying deceptive messaging. Truthy, for example, is intended to spot social-media campaigns that are orchestrated by an individual or organization, but purport to be a spontaneous expression by a large group of independent individuals [42]. Another research team has focused on detecting deceptive "opinion spam", favorable reviews written by individuals in exchange for compensation [40]. Videolyzer allows users to analyze online political videos and rate them in terms of accuracy and bias [11] and SRSR (pronounced "sourcer") aims to help journalists identify trustworthy content amidst the flood of information produced by social media [10].

Another vein of research in this area involves creating systems that can enhance individuals' understanding of a contentious issue by mapping relationships among competing claims and corresponding evidence using a mix of manual and automated processing [24, 45] (e.g., Debatepedia, DebateGraph, Cohere, considerate, and many others). Some of these systems have an explicitly evaluative component, and are intended to help users systematically review relevant data so that they might assess a claim or prediction [35] (e.g., Statement Map, Competing Hypotheses). Other systems are designed to facilitate citizens' exposure to a diverse range of information and opinions. Prototypes strive to present readers with news articles containing multiple viewpoints [41], or to tailor news aggregators in order to encourage a diverse diet of opinions [34].

Finally, there are a handful of systems that attempt to correct online misinformation at the point of contact. Perhaps the most well known of these is Dispute Finder [13, 14]. The goal of this system is to highlight inaccurate phrases on a webpage as the page is displayed. It accomplishes this with a browser plug-in that executes a simple text entailment algorithm, comparing the content of page to be displayed to a cached database of previously identified falsehoods. The corpus of claims is constructed by crawling Snopes and Politifact and through manual additions by system users. The result is a tool that allows users easy access to fact-checking information from any webpage that includes one of the disputed claims. In 2009 and 2010, a commercial service with similar objectives was also being developed. Called DotSpots, the system allowed users to annotate text on one webpage and would automatically display the annotation on other pages containing similar text. Crowdsourced fact checking was not the sole goal of the system, but the company's slogan "spot the truth, connect the dots!" suggests that this was a prominent consideration. This service, however, failed to achieve a critical mass of users, and development ended in late 2010. Nevertheless, interest in real-time corrections continues unabated, as exemplified most recently in Hypothes.is. Like its predecessors, the project aims to discourage the flow of inaccurate information, this time by creating a distributed platform for textual annotation paired with a reputation system intended to ensure the annotations are of high quality.

In sum, there is a growing collection of tools that can be used to identify, track, analyze, and potentially correct misinformation online, and several systems aim to present these corrections in real time. Which brings us to the question driving this research, namely, how effective is a real-time fact-checking approach? Research in psychology offers some insights into how we answer this question.

Political fact-checking psychology

Correcting misinformation as it is presented assumes that political learning is a simple operation, with individuals retaining information that is supported by facts and discounting or rejecting unsupported claims and falsehoods. On this view, correcting a misperception is no different than learning about a change in a weather forecast: you were expecting sun, but now the forecast calls for rain. In this scenario, updating your beliefs is straightforward and unambiguous. You were not invested in the old forecast, the meteorologist has nothing to gain from lying, and the cost of being wrong is relatively low.

Political learning, however, is rarely so simple. There are important differences between a weather forecast and a contentious political issue such as climate change. In contrast to a forecast, people are more likely to have invested effort into reaching an opinion about climate change. Furthermore, experts making the claims—those with the data and the corresponding analytic tools—have a stake in what people believe, and the costs of being wrong about the issue are high. Given this, it should be unsurprising that people cautiously approach new information on contentious topics, especially information that runs counter to their beliefs. That people argue against counterattitudinal evidence while readily accepting proattitudinal evidence is undisputed [29, 31, 51]. Several different factors may contribute to this behavior. It could be at least partially an artifact of Bayesian learning, whereby the same information has different consequences for belief depending on the individual's prior knowledge and confidence [7, 19]. But there is also compelling evidence of an affective component, whereby individuals are motivated to defend their position beyond the point of reason [43]. The result is that beliefs can diverge in response to corrections; in some extreme cases, individuals may even become more accepting of inaccurate information [39].

In short, there is significant evidence that people do not simply "learn" from fact-checking messages. Instead, individuals assess new information before storing it in memory, evaluating whether the evidence is sufficiently persuasive to merit updating their beliefs and adjusting the magnitude of the update based on both the strength of their prior convictions and of the novel evidence [54]. This is neither surprising nor fundamentally problematic—to unquestioningly accept every new claim encountered would be naïve—but the practice has important implications for how people respond to attempts to correct misperceptions. Specifically, it suggests that it may be most appropriate to view factual corrections as a form of persuasion or strategic communication [4]. After all, these messages are intended to convince the reader that they represent the truth and that claims to the contrary are false. This has important consequences for how we understand the response that corrective messages elicit.

There are some reasons to expect corrections presented at the time of exposure to perform better than those presented later, even if people vet counterattitudinal information. People are not "ambulatory encyclopedias", memorizing every fact they encounter; instead they rely on shortcuts to arrive at their decisions [30]. One important shortcut involves maintaining easily recalled summary judgments of attitudes and beliefs, while discarding much of the evidence on which the judgments are based [3, 54]. If the evidence contained in a correction is convincing enough that recipients are compelled to accept it—even if it means that they only reject parts of the misinformation—then an immediate correction should reduce the risk of individuals updating stored attitudes based on a falsehood. On this view, flagging inaccuracies before they can sway attitude is potentially useful, and real-time corrections should generally outperform delayed corrections.

What, then, is the risk? The key question is whether a correction embedded in an inaccurate message generates *more* resistance than a correction presented at a later time. There are a few reasons to think that it could. First, real-time corrections may produce heightened counterargument. Corrections presented at the time of exposure are more confrontational because they draw attention to points of controversy, directly challenging claims that some readers are inclined to believe. This can create an affective response—e.g., anger and defensiveness at being told one's views are based on lies—and can be more threatening to those inclined to believe the falsehood than are corrections not attached to specific claims. Research has shown that people generate more counterarguments to ego-threatening messages than non-threatening messages [26], and that such counterargument makes attitude-consistent evidence more accessible [6, 15, 39]. As a consequence, threat-inducing corrections are more likely to be overwhelmed by arguments in favor of a misperception than less threatening messages.

Second, correcting inaccuracies at the time of exposure could encourage users to question the credibility of the message. Real-time corrections explicitly pit two claims against one another: the misinformation and its correction. Forced to weigh these competing claims, individuals are likely to regard the attitude-consistent claims as more believable [29]. Furthermore, there is evidence that this tendency becomes stronger when the individual feels more threatened. Source derogation, for example asserting that a source is biased or that it lacks relevant expertise, is more prevalent when individuals feel their position is in jeopardy [26], and this reduces the likelihood that an individual will act in accordance with a corrective message [8].

In sum, although we do anticipate that real-time corrections should be more effective than delayed corrections overall, they may be less effective among those who are predisposed to believe the misinformation. For these individuals, corrections applied directly to an inaccurate

Figure 1. Partial screenshot showing an embedded correction.

claim could induce greater counterargument and could lead the recipient to question the message's credibility.

EXPERIMENTAL STUDY

We conducted a between-participants experiment to examine how real-time delivery of fact checking information influences recipients' subsequent beliefs compared to other strategies. The topic of the information presented in this study was electronic health records (EHRs), an issue that had received only modest news coverage at the time of data collection (May 2011). We utilized an opt-in online panel administered by Survey Sampling International to recruit a demographically diverse sample of U.S.-based participants (N = 574). The sample is 49% male, has an average age of 45.8 years (SD = 15.8), and is racially diverse (86.9% While, 6.8% Black, 6.3% other). Participants also had a range of party affiliations (25.4% Republican, 34.7% Democrat, 28.1% Independent, 11.9% other) and of ideologies (28.4% Liberal, 35.0% Moderate, 36.6% Conservative).

Procedure

The experiment compared participants' beliefs across three conditions. In all conditions, we began by asking participants to tell us how much they knew about five contemporary policy issues, including electronic health records. Familiarity was measured on a seven-point scale, anchored by "unfamiliar with the issue" (coded as 1) and "know a great deal about the issue" (M = 4.2, SD = 1.7). We also asked participants to indicate their attitude toward the same five issues on a seven-point scale anchored by "extremely negative" (coded as 1) to "extremely positive" (M = 4.8, SD = 1.6).

Next, we asked participants to read a 443-word "news article" written by a journalist with guidance from the research team that provided a brief introduction to EHRs, describing the technology, its objectives, and current deployment levels. This information was gleaned from contemporary news stories and government sources, and was accurate to the best of our knowledge.[1] Participants were required to spend at least one minute viewing the story, though many spent substantially longer (M = 112s, SD = 78s).

The three conditions diverged at this point, varying how inaccurate information was subsequently presented and corrected. In the first condition, the *delayed correction* (n = 191), participants were next shown a 367-word message that contains a number of factual errors, purportedly copied from "a widely read political blog". The errors, which we inserted intentionally, include several false statements about who is allowed to access EHRs. For instance, the message claims that hospital administrators, health insurance companies, employers, and government officials have unrestricted access to personal health information. As before, participants were required to spend at least one minute viewing the story, though the average participant spent more than the minimum (M = 93s, SD = 51s).

Before presenting a correction, participants in the delayed-correction condition were asked to complete a three-minute image-comparison task. The directions stated that this would allow researchers to understand how the individual processes images, but its true function was to serve as a distractor task, clearing working memory prior to introducing the correction. Participants were presented with a pair of nearly identical images and had one minute to count observed differences before reporting their results (M = 4.58, SD = 2.17, range 0-11). Participants were then informed that there were in fact 13 differences and were encouraged to be as accurate as possible in subsequent comparisons. The comparison task was repeated twice more without feedback.

After completing the distractor task, participants were presented with a 378-word correction attributed to FactCheck.org, an award-winning non-partisan news service. The correction addressed each of the inaccuracies included in the previous message, noting for example that there are clear policies restricting access to patient health information to those involved in a patient's care. Most participants spent more than the required minute reading this article (M = 105s, SD = 63s).

In the second condition, the *immediate correction* (n = 182), participants were presented with an annotated version of the "blog post" described above (see Figure 1). The directions explained that, "A third-party fact-checking service has reviewed this blog post and concluded that it contains factual errors. Inaccurate statements are italicized, enclosed in [square brackets] and displayed in red. Please see the fact-checking article at the bottom of this page for more detailed information." All false information was marked in the body of the message accordingly. Below this, the fact-checking message used in the first condition was presented in its entirety. Participants were required to spend at least two minutes reading the corrected document, but most spent considerably longer (M = 181s, SD = 87s).

[1] Stimuli are available from the first author upon request.

The visual flagging of false claims in this condition is similar, but not identical, to Dispute Finder's interface. In Dispute Finder, the claim was highlighted in red; in this study it was printed in dark-red italicized text and enclosed in square brackets. We selected this presentation style to ensure that individuals could spot inaccuracies even if they were not easily able to see red highlighting. Also, Dispute Finder corrections appeared in a popup if the user clicked on the highlighted snippet; in contrast, corrections in this condition were always present at the bottom of the page. We know from prior online experiments that participants sometimes have difficulty managing popups, and we did not want this to be an obstacle to successful completion of the study.

In the *control condition* (n = 201), participants were only presented with the inaccurate message during the study; the correction was presented after the study was complete, during debriefing.

The study concluded with a brief questionnaire, beginning with a series of standard psychological measures. One of these was a "memory tally", which asked participants to list everything they learned about EHRs from the reading (up to ten items, M = 3.5, SD = 2.7). Next, participants were asked to indicate their feelings about each recalled item on a seven-point scale anchored by "extremely negative" (coded as 1) and "extremely positive". We then counted the number of negative items (those scored below the scale midpoint) that came to mind (M = 1.3, SD = 1.7).

Accuracy was measured by asking participants to indicate how easy or difficult it will be for each of several groups (doctors, employers, government officials, hospital administrators, insurers, pharmaceutical companies, and medical staff—listed in random order) to access EHRs using a seven-point scale anchored by "very easy" (coded as 1) and "very difficult". Responses most consistent with the fact-checking document would describe doctors and medical staff as having very easy access and everyone else as having very difficult access. Thus, the items corresponding to the first two groups were reverse coded and the seven items were summed to create an accuracy measure (α = .75, M = 28.8, SD = 8.2, range 13-49).

Finally, the questionnaire asked participants to assess the credibility of the fact-checking message by answering three questions: "How successful was the fact checking article at discrediting the claim that Electronic Health Records will allow limitless access to patient health information?", "How persuasive was the evidence given in the fact checking article that Electronic Health Records do not pose a privacy threat?", and "How credible was the fact checking article's presentation of information about Electronic Health Records?". The questions used a seven-point response scale, with higher scores corresponding to higher credibility (α = .88, M = 14.3, SD = 4.5).

Results

First we confirm that corrections can be effective, even on politically charged topics. Results indicate that individuals exposed to a fact-checking message hold more accurate beliefs than those who are not exposed. To see this, we constructed a linear regression model predicting belief accuracy by condition, treating the control as the reference category. Four cases in which the participant did not answer all the belief questions are omitted, leaving an n of 570. Interpreting model coefficients (not shown, but see Figure 2), we find that compared to the control condition participants in both the delayed-correction condition (*diff*=3.3 points, $p < .001$) and immediate-correction condition (*diff*=5.2 points, $p < .001$) score significantly higher on the accuracy measure. This establishes that these beliefs, which are colored by political interests, can change in light of new information.

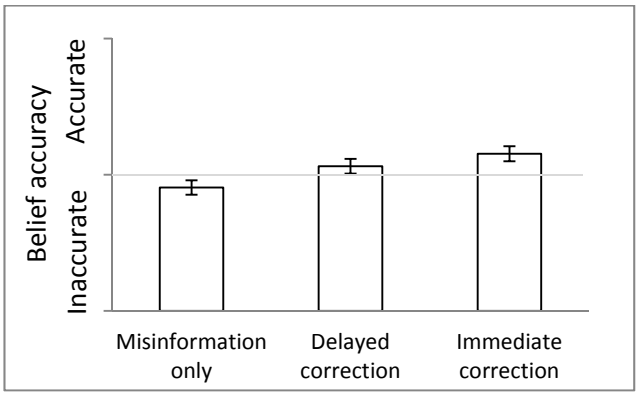

Figure 2. Predicted belief accuracy by condition based on linear regression coefficients; 95% confidence intervals shown. The grey horizontal line is the accuracy scale midpoint (28).

We also posit that immediate corrections are more effective than corrections presented after a delay, as implicitly assumed by many designers who have embraced the real-time approach. The data support this prediction as well. Visually, we see that the bar representing beliefs in the immediate-correction condition is taller than that of the delayed-correction condition. A Wald test confirms that this difference is significant. The coefficient on the immediate-correction condition is significantly larger than that of the delayed correction, $F(1, 567) = 5.31$, $p < .05$.

Recall, however, that we also anticipated real-time corrections having some harmful effects. More important than the modest improvements in accuracy associated with an immediate correction is the possibility that this approach might amplify the influence of attitude-based bias. The data suggest that this is the case. Adding interaction terms between condition and participants' issue favorability (mean centered) to the previously described regression allows us to test these relationships (see Table 1).

As one would expect, an individual's attitude toward Electronic Health Records prior to viewing the stimuli has a strong main effect on accuracy: the more (less) positively

the individual feels about the issue, the more (less) accurate his or her beliefs. When misinformation is corrected immediately, this difference is more pronounced: correcting misinformation at the time of exposure is more effective for issue supporters, but less effective among opponents. There is, however, no evidence that issue favorability has a moderating effect in the delayed correction: the correction's influence is the same regardless of the participants' issue favorability. Figure 3 illustrates these relationships, highlighting the fact that the effectiveness of a real-time correction is due to its performance among those most inclined to reject the misinformation. Among those who oppose EHRs, the effect of the immediate correction on beliefs is statistically comparable to no correction at all.

	B	***SE***
Delayed correction [a]	3.50***	(0.78)
Immediate correction [a]	5.24***	(0.79)
Issue favorability	0.99**	(0.36)
Delayed * favorability	0.03	(0.52)
Immediate * favorability	1.07*	(0.51)
Intercept	25.92***	(0.55)

Table 1 . Linear regression predicting belief accuracy. Note $N = 570$, $R^2 = 0.14$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ Favorability is mean-centered. ([a]) Misinformation-only condition is reference category.

We identified two mechanisms that might help explain the uniquely biased processing of real-time corrections. First, corrections that are presented alongside a false claim might cause issue opponents to engage in more vigorous counterargument than corrections presented later. A linear regression predicting the number of counterarguments listed during the memory-recall task, however, offers no support for this prediction (Table not shown). The coefficients on both the immediate and the delayed correction conditions are negative—$B = -.47$, $p < .01$ for the delayed correction and $B = -.72$, $p < .001$ for the immediate correction—and they are not significantly different from one another, $F(1, 571) = 2.42$, $p = .12$. In other words, individuals volunteered *fewer* negative thoughts about EHRs when exposed to a correction than not, regardless of when the correction was presented. Furthermore, issue favorability has no influence on this relationship. The interactions between issue favorability and each of the two correction conditions are non-significant. Participants in these two conditions have comparable numbers of negative thoughts in response to the correction, regardless of their prior attitude toward EHRs. Hence, there is no evidence that real-time corrections provoke heightened counterargument, even among issue opponents.

The second mechanism we identified to help explain biased responses to real-time corrections was the idea that

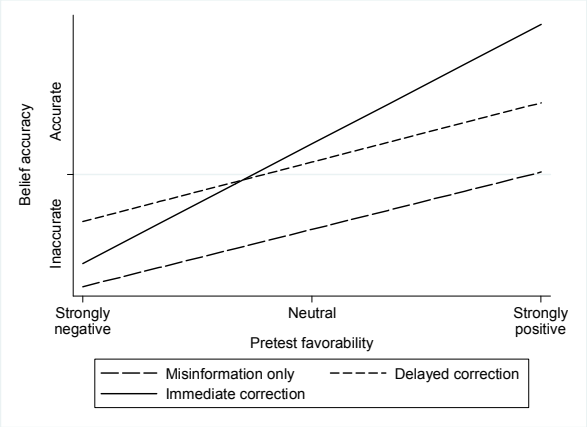

Figure 3. Predicting belief accuracy by issue favorability for each condition based on linear regression coefficients.

embedded corrections effectively pit two claims against one another, perhaps causing users to place greater focus on the credibility of the competing messages. Since credibility assessments tend to be biased by prior attitudes, focusing on them should lead EHR opponents to question the fact-checking message's credibility. The data support this explanation, as evidenced by a linear regression model predicting perceived credibility of the correction (see Table 2 and Figure 4). Consistent with prior research, we see that the more favorably an individual felt about EHRs, the more credible the correction was perceived to be. The significant interaction term indicates that this relationship is stronger in the immediate-correction condition than in the delayed-correction condition, suggesting that attitude-based biases play a bigger role when corrections are presented in real time.

	B	***SE***
Immediate correction [a]	0.79	(0.43)
Issue favorability	0.85***	(0.20)
Immediate * favorability	0.57*	(0.28)
Intercept	13.94***	(0.30)

Table 2 . Linear regression predicting message credibility. Note $N = 372$, $R^2 = 0.17$, * $p < 0.05$, *** $p < 0.001$ Favorability is mean-centered. ([a]) Delay is reference category.

Finally, we considered whether issue favorability influenced belief accuracy indirectly by shaping participants' trust of the corrective message. A mediation test using boot-strapped confidence intervals confirms this, demonstrating that the influence of issue favorability on belief accuracy is mediated by the perceived credibility of the message [21]. This holds when the correction is immediate, with a 95% confidence interval for the mediated effect between .94 and 1.97; and when it is delayed, with a

95% confidence interval between .42 and 1.35. In other words, individuals who are more favorably inclined toward EHRs view the corrections as more credible, which in turn promotes more accurate beliefs.

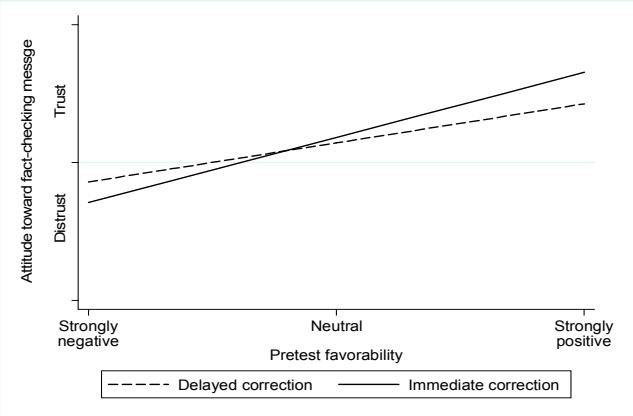

Figure 4 . Predicting credibility of correction by issue favorability for each correction condition based on linear regression coefficients.

DISCUSSION

Despite the obvious appeal of providing immediate corrections to false information online, this approach has some negative consequences that need to be addressed. Annotating a misleading message by highlighting inaccuracies and embedding fact-checking information accentuates individuals' tendency to view these corrections through an attitudinally biased lens. We should be careful not to exaggerate the significance of this behavior: real-time corrections are often as effective as traditional post-exposure correction strategies. The problem is that these techniques actually increase resistance to the correction among those whose attitudes are most strongly supported by the misperception. Thus, for example we would anticipate that systems like Dispute Finder would do little to change the beliefs of the roughly one in six Americans who, despite exhaustive news coverage and fact checking, continue to question whether President Obama was born in the U.S. [2] or whether vaccines are safe [20]. New approaches may be required when designing automated systems to promote more accurate beliefs among those whose prior attitudes leave them predisposed to hold a misperception.

Implications for Practice

We believe that the key theoretical insight for guiding future design work is our argument that people respond to fact-checking messages in ways that are comparable to their response to propaganda and persuasion. Indeed, critics at both ends of the political spectrum have labeled fact checking organizations as partisan [22, 53]. This comparison implies a more complicated view of political learning than assumed in prior design work, but it also suggests a variety of strategies from which system builders may be able to learn. Scholars who study health-behavior modification campaigns (e.g., anti-smoking campaigns) and those who work in science communication (e.g., efforts to increase public understanding of global climate change) have been grappling with similar challenges for a long time, and some of the approaches they have developed may translate. We highlight a few strategies that we believe are particularly promising.

Recommendation 1. One strategy focuses on the source of the correction. It is hardly surprising that individuals trust some sources more than others [33], and that trust in expertise varies significantly based on the attributes of both the expert and the individuals [5, 37]. For example, independent experts are generally more influential than industry representatives, but there are many individuals who tend to distrust both groups. Research suggests that likeable sources and sources that share characteristics with the message recipient are less prone to derogation [8]. Criticisms from unexpected sources are also uniquely persuasive, as when a politician criticizes one of his or her ideological allies [1].

This suggests that users may be more willing to accept corrective messages from sources that are ideologically similar, or that they choose for themselves. To accomplish the former, a system might offer corrective messages derived from and attributed to frequently used sources, or sources that the user regularly recommends to others (via Facebook Likes or Tweets, for example). Alternatively, designers might allow users to explicitly tailor which sources the correction system utilizes. An architecture that allows multiple independent groups to create their own databases of disputed claims and affords users the opportunity to opt in to one or more of these differing collections could be beneficial.

A question raised by this approach is who decides what the truth is. The political world is complex, and there are legitimate alternative interpretations of many situations. At the same time, however, there are instances when the preponderance of evidence supports one factual claim over another (e.g., President Obama was born in the U.S.; the 9/11 attacks were not a Bush-administration conspiracy). Fortunately, experts who disagree with one another ideologically often agree about the facts, though they may differ in how they present them. Thus, allowing users to select which sources they trust does not mean giving up on the question of truth. Furthermore, although there is a risk that individuals might seek out "fact-checking" sources that affirm information consistent with their ideological beliefs without regard to the evidence, while avoiding other sources, there is growing evidence that such an aversion is relatively rare [17] and a competitive media environment can help to guard against outlet bias by penalizing the reputation of sources that are consistently inaccurate [18].

Recommendation 2. Another strategy concerns how the message is framed. For example, a correction might highlight the *risks* associated with holding a misperception,

the *benefits* of a more accurate belief, the *moral obligation* to weigh the evidence fairly, etc. Different frames have been shown to resonate with different audiences, and the better aligned the frame is with the recipients' ideology, the more likely the message is to be accepted [32]. For example, conservatives tend to be uniquely responsive to information framed in terms of loss, while liberals respond better to benefit frames [27]. The power of how a correction is framed should not be underestimated: some research suggests that the influence of factual content is almost entirely eclipsed by its framing. Experimental studies have shown that compared to how a message is framed, factual information has comparatively little sway on opinion formation and is more prone to being viewed as biased [12].

A fact-checking system might tailor its presentation of corrections by combining user profiles, algorithms for guessing frame preferences based on these profiles, and databases of corrections framed in a variety of ways. For example, users who consistently consume a diverse mix of ideologically oriented outlets might be uniquely responsive to a message framed in terms of the need to weigh competing evidence. Individuals who have been identified as having a high risk aversion, either through self-reports or, more likely, implicit behavior measures, would instead see corrections that emphasis the risks associated with inaccurate beliefs on the topic.

It would be crucial in such a system, however, to ensure that the factual information is consistent across the various frames least it become a propaganda tool. Furthermore, the creation of such a system poses considerable threat to user privacy. Although many search services use behavior logs to inform future results, implementing this recommendation requires careful consideration. It could, for example, be problematic to maintain a remote database of user frame-preference profiles without a mechanism for protecting the identity of those users lest the information be abused. Framing strategies are also a double-edged sword: the system imagined here could also be uniquely effective for introducing inaccuracies. Despite these risks, this approach could yield significant benefits. If we can present accurate information in ways that allow individuals to be more receptive to it while protecting user privacy and guarding against system abuse, we may be able to significantly reduce misperceptions.

Recommendation 3. The last strategy is suggested by our growing understanding of the psychological mechanisms that motivate biased processing. As noted in our initial theorizing, ego-threats have been shown to promote counterargument and source derogation. Other work has shown that self-affirmation—such as reflecting on one's positive attributes—prior to exposure to a counterattitudinal message can work in the opposite direction and reduce bias [9, 52]. Furthermore, inducing positive feelings toward the self is relatively straightforward [36]. This suggests that fact-checking systems may be more effective if corrections are timed to follow self-affirming experiences, either naturally occurring or purposefully constructed.

Fact-checking systems could monitor user-behavior, delivering corrective information after individuals have consumed content that is self-affirming, such as news stories that reinforce their political values, personal beliefs, or personal contributions to the world. Alternatively, these systems could couch the corrections in a positive message, for example acknowledging the sources that user has already consulted and complimenting the user's prior efforts to consider alternative perspectives.

It is important that the system not be perceived as manipulative or patronizing. Instead, the affirmation should be a legitimate positive experience that can help to counteract the unease that typically accompanies being corrected. It should be noted, however, that recent experimental work has raised questions about this approach. A series of studies found that although self-affirmation led participants to express more accurate beliefs, it did not augment the effects of corrective information: accuracy improvements occurred regardless of whether a correction was presented or not [38]. Nevertheless, given the somewhat inconsistent evidence on this topic, more research is merited.

Limitations

This study has several limitations. First, these results are based on a single exposure to a correction. It is possible that the effects may be more promising when users are presented with a regular stream of factual corrections, as research has shown that partisan biases do exhibit a tipping-point effect, whereby new information eventually overcomes prior predispositions [44]. Such a study could be accomplished in the lab, with an information board design providing repeated exposures over an extended period, or in the wild, through user studies (we return to this idea below).

Second, the study was conducted in the context of a single issue: electronic health records. Although the misinformation was designed to elicit negative reactions from across the political spectrum (e.g., concerns about big government for conservatives and about big business for liberals), Democrats were initially more favorable to the issue than Republicans, 5.2 versus 4.4, t (434) = 5.3, $p <$ .001. Thus, it is possible that the effects seen here are limited to a particular demographic group. More research is needed, but we believe that the problems demonstrated are likely to apply across party lines, as shown in other research [39].

Third, differences between the immediate and delayed corrections could be due (at least in part) to the distractor task itself, which was only present in the latter condition. The distractor could have two possible effects. It could increase cognitive load of the task, which would artificially

reduce participants' performance in the delayed condition. If this were the case, however, it would only strengthen our findings: including a distractor in both conditions would increase the relative advantage of the delayed correction over the immediate correction. The other possible effect is grounded in participants' performance on the distractor task. Participants who did well on the task, correctly identifying the number of differences between the pictures shown in the first image pair, might experience a boost in self-confidence, akin to the self-affirmation manipulations described above, and this could lead them to be more receptive to corrections. The data, however, do not support this interpretation: the correlation between the accuracy of identified differences in the distractor and the accuracy of factual beliefs was non-significant (and very small).

Fourth, it is plausible that the immediate-correction condition, in which misinformation and its rebuttal are presented simultaneously, could influence how corrections are encoded by encouraging participants to move back and forth between the two types of content. This could have one of two consequences. On one hand, the effort invested in encoding the information might lead to better memory and more robust long-terms effects of the immediate correction despite the somewhat discouraging short-term effects. On the other hand, presenting a claim and counter-claim side-by-side could prompt more effortful processing, especially among those invested in the original claim. Recent work in social psychology suggests that the more effort required to understand a statement, the less likely it is to be believed [46]. Thus, if the immediate correction produces more effortful processing than the delayed correction, it will be less persuasive. This study did not capture participants' scrolling behavior, so we cannot rule this possibility out entirely. We did, however, collect timing data. If we assume that effortful cross-content comparison is time consuming, than the preceding arguments suggest that the cumulative time participants spent on misinformation and its correction should differ by condition. Instead, we find that total viewing time in the immediate and delayed conditions is not statistically different (immediate = 182s, delayed = 191s, p = .07). Hence, there is no evidence that the effort participants' exerted to process the messages differed across conditions.

Finally, it is possible that these results are specific to the visual presentation style employed here. This seems unlikely: we can think of no theoretical reason to expect pop-ups to be more trusted or less threatening than on-screen corrections, or for red highlighting to induce less bias than red italicized text. Nevertheless, this is an empirical question that would be effectively answered with a subsequent study.

We believe that we offer reasonable evidence against each of the rival explanation identified here. Nevertheless, a real-world test of competing designs could provide useful additional evidence. Comparing the belief accuracy of individuals using a system that provides immediate corrections (perhaps built on top of the foundation offered by Dispute Finder, or promised by Hypothes.is) to one that presents a cumulative summary of corrective information after a delay would be informative. Such a test could be conducted over a period of weeks or months, shedding light on long-term dynamics. This approach would also offer greater ecological validity, helping to ensure that the effects are not a product of the experimental design.

CONCLUSION

Fostering a better informed citizenry is an admirable goal with many potential benefits, including better policy decisions, better health choices, and more. Using computer software to augment humans' ability to sift through the vast stores of online information, distinguishing fact from fiction, is a potentially crucial tool for accomplishing this. People do not have time to systematically evaluate every claim they encounter, and the value of helping them achieve an understanding that reflects the best evidence on any issue is undisputed. This paper demonstrates, however, that the complexity of building software is only part of the challenge.

Providing factual information is a necessary, but not sufficient, condition for facilitating learning, especially around contentious issues and disputed facts. As highlighted by this study, individuals are influenced by a variety of biases that can lead them to reject carefully documented evidence, and correcting misinformation at its source can actually augment the effects of these biases. Our goal is not to discourage future work in this area, but to suggest a variety of correction-presentation strategies the designers might use to help overcome these biases.

ACKNOWLEDGEMENTS

The authors also gratefully acknowledge Emily Lynch for her assistance with the design of stimuli and manipulations, and Cliff Lampe, Paul Resnick, Scott Robertson and members of the Ohio State Communication, Opinion, and Political Studies group for their thoughtful comments and suggestions. This work was funded in part by the U.S. National Science Foundation under Grant No. IIS-1149599.

REFERENCES

1. Baum, M. and Groeling, T. Shot by the Messenger: Partisan Cues and Public Opinion Regarding National Security and War. *Political Behavior*, *31*, 2 (2009), 157-186.
2. Berinsky, A.J. The Birthers Are Back. http://today.yougov.com/news/2012/02/03/birthers-are-back/.
3. Bizer, G.Y., Tormala, Z.L., Rucker, D.D. and Petty, R.E. Memory-based versus on-line processing: Implications for attitude strength. *Journal of Experimental Social Psychology*, *42*, 5 (2006), 646-653.

4. Bordia, P., DiFonzo, N., Haines, R. and Chaseling, E. Rumors Denials as Persuasive Messages: Effects of Personal Relevance, Source, and Message Characteristics. *Journal of applied social psychology*, *35*, 6 (2005), 1301-1331.
5. Brossard, D. and Nisbet, M.C. Deference to Scientific Authority Among a Low Information Public: Understanding U.S. Opinion on Agricultural Biotechnology. *International Journal of Public Opinion Research*, *19*, 1 (2007), 24-52.
6. Bullock, J.G. Experiments on partisanship and public opinion: Party cues, false beliefs, and Bayesian updating *Ph.D. Dissertation, Stanford University*, Stanford, CA, 2007.
7. Bullock, J.G. Partisan Bias and the Bayesian Ideal in the Study of Public Opinion. *Journal of Politics*, *71*, 3 (2009), 1109-1124.
8. Byrne, S. and Hart, P.S. The 'boomerang' effect: A synthesis of findings and a preliminary theoretical framework. *Communication Yearbook*, *33* (2009), 3-37.
9. Cohen, G.L., Aronson, J. and Steele, C.M. When Beliefs Yield to Evidence: Reducing Biased Evaluation by Affirming the Self. *Personality and Social Psychology Bulletin*, *26*, 9 (2000), 1151-1164.
10. Diakopoulus, N., De Choudhury, M. and Naaman, M., Finding and Assessing Social Media Information Sources in the Context of Journalism. In *Proc. Conference on Human Factors in Computing Systems (CHI)*, ACM (2012).
11. Diakopoulus, N., Goldenberg, S. and Essa, I., Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists In *Proc. Conference on Human Factors in Computing Systems (CHI)*, ACM Press (2009).
12. Druckman, J.N. and Bolsen, T. Framing, Motivated Reasoning, and Opinions About Emergent Technologies. *Journal of Communication*, *61*, 4 (2011), 659-688.
13. Ennals, R., Byler, D., Agosta, J.M. and Rosario, B., What is disputed on the web? In *Proceedings of the 4th workshop on Information credibility*, ACM (2010), 67-74.
14. Ennals, R., Trushkowsky, B. and Agosta, J.M., Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*, ACM (2010), 341-350.
15. Freedman, J.L. and Sears, D.O. Warning, distraction, and resistance to influence. *Journal of Personality and Social Psychology*, *1*, 3 (1965), 262-266.
16. Garrett, R.K. Troubling consequences of online political rumoring. *Human Communication Research*, *37*, 2 (2011), 255-274.
17. Garrett, R.K., Carnahan, D. and Lynch, E. A turn toward avoidance? Selective exposure to online political information, 2004-2008. *Political Behavior* (2011), 1-22.
18. Gentzkow, M. and Shapiro, J.M. Media bias and reputation. *Journal of Political Economy*, *114* (2006), 280-316.
19. Gerber, A. and Green, D. Misperceptions about perceptual bias. *Annual Review of Political Science*, *2*, 1 (1999), 189-210.
20. Harris Interactive/HealthDay. Vaccine-Autism Link: Sound Science or Fraud? . http://www.harrisinteractive.com/newsroom/pressreleases/tabid/446/mid/1506/articleid/674/ctl/readcustom%20default/default.aspx.
21. Hayes, A.F., Preacher, K.J. and Myers, T.A. Mediation and the Estimation of Indirect Effects in Political Communication Research. in Bucy, E.P. and Holbert, R.L. eds. *Sourcebook for Political Communication Research: Methods, Measures, and Analytical Techniques*, Routledge, New York, 2011, 434-465.
22. Hemingway, M. Lies, damned lies, and 'fact checking' *The Weekly Standard*, 2011.
23. Katz, J.E. Struggle in cyberspace: fact and friction in the World Wide Web. *Annals of the American Academy of Political and Social Science*, *560*, 1 (1998), 194-199.
24. Kriplean, T., Morgan, J.T., Freelon, D., Borning, A. and Bennett, L., ConsiderIt: improving structured public deliberation. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, ACM (2011), 1831-1836.
25. Kuklinski, J.H., Quirk, P.J., Jerit, J., Schwieder, D. and Rich, R.F. Misinformation and the Currency of Democratic Citizenship. *The Journal of Politics*, *62*, 03 (2000), 790-816.
26. Lapinski, M.K. and Boster, F.J. Modeling the ego-defensive function of attitudes. *Communication Monographs*, *68*, 3 (2001), 314-324.
27. Lavine, H., Burgess, D., Snyder, M., Transue, J., Sullivan, J.L., Haney, B. and Wagner, S.H. Threat, Authoritarianism, and Voting: An Investigation of Personality and Persuasion. *Personality and Social Psychology Bulletin*, *25*, 3 (1999), 337-347.
28. Leskovec, J., Backstrom, L. and Kleinberg, J., Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM Press (2009), 497-506.
29. Lord, C.G., Ross, L. and Lepper, M.R. Biased assimilation and attitude polarization: the effects of prior theories on subsequently considered evidence.

Journal of Personality and Social Psychology, *37*, 11 (1979), 2098-2109.

30. Lupia, A. and McCubbins, M.D. *The Democratic Dilemma*. Cambridge University Press, Cambridge, 1998.
31. MacCoun, R.J. and Paletz, S. Citizens' Perceptions of Ideological Bias in Research on Public Policy Controversies. *Political Psychology*, *30*, 1 (2009), 43-65.
32. Maibach, E.W., Roser-Renouf, C. and Leiserowitz, A. Communication and Marketing As Climate Change–Intervention Assets: A Public Health Perspective. *American Journal of Preventive Medicine*, *35*, 5 (2008), 488-500.
33. Moser, S.C. Communicating climate change: history, challenges, process and future directions. *Wiley Interdisciplinary Reviews: Climate Change*, *1*, 1 (2010), 31-53.
34. Munson, S.A. and Resnick, P., Presenting Diverse Political Opinions: How and How Much. In *CHI 2010* (2010).
35. Murakami, K., Nichols, E., Mizuno, J., Watanabe, Y., Masuda, S., Goto, H., Ohki, M., Sao, C., Matsuyoshi, S., Inui, K. and Matsumoto, Y., Statement map: reducing web information credibility noise through opinion classification. In *Proceedings of the fourth workshop on Analytics for noisy unstructured text data*, ACM Press (2010), 59-66.
36. Napper, L., Harris, P. and Epton, T. Developing and Testing a Self-affirmation Manipulation. *Self and Identity*, *8*, 1 (2009), 45-62.
37. Nisbet, M.C. and Scheufele, D.A. What's next for science communication? Promising directions and lingering distractions. *American Journal of Botany*, *96*, 10 (2009), 1767-1778.
38. Nyhan, B. and Reifler, J. Opening the Political Mind?: The effects of self-affirmation and graphical information on factual misperceptions, 2011.
39. Nyhan, B. and Reifler, J. When Corrections Fail: The persistence of political misperceptions. *Political Behavior*, *32*, 2 (2010), 303-330.
40. Ott, M., Choi, Y., Cardie, C. and Hancock, J.T., Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, Association for Computational Linguistics (2011), 309-319.
41. Park, S., Kang, S., Chung, S. and Song, J., NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the 27th international conference on Human factors in computing systems*, ACM (2009), 443-452.
42. Ratkiewicz, J., Conover, M., Meiss, M., Gonçalves, B., Patil, S., Flammini, A. and Menczer, F., Detecting and Tracking the Spread of Astroturf Memes in Microblog Streams. In *Proceedings of the 20th international conference companion on World wide web* (2010), 249-252.
43. Redlawsk, D.P. Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *Journal of Politics*, *64*, 4 (2002), 1021-1044.
44. Redlawsk, D.P., Civettini, A.J.W. and Emmerson, K.M. The Affective Tipping Point: Do Motivated Reasoners Ever "Get It"? *Political Psychology*, *31*, 4 (2010), 563-593.
45. Schneider, J., Groza, T. and Passant, A. A Review of Argumentation for the Social Semantic Web. *Sematic Web journal*, *4*, 3 (to appear 2013).
46. Schwarz, N., Sanna, L.J., Skurnik, I. and Yoon, C. Metacognitive Experiences and the Intricacies of Setting People Straight: Implications for Debiasing and Public Information Campaigns. in Zanna, M.P. ed. *Advances in experimental social psychology*, Academic Press, 2007, 127-161.
47. Seifert, C.M. The continued influence of misinformation in memory: What makes a correction effective? in Brian, H.R. ed. *Psychology of Learning and Motivation*, Academic Press, 2002, 265-292.
48. Simmons, M.P., Adamic, L. and Adar, E., Memes Online: Extracted, Subtracted, Injected, and Recollected. In *Fifth International AAAI Conference on Weblogs and Social Media* (2011).
49. Stempel, C., Hargrove, T. and Stempel III, G.H. Media use, social structure, and belief in 9/11 conspiracy theories. *Journalism & Mass Communication Quarterly*, *84*, 2 (2007), 353-372.
50. Sunstein, C.R. *On rumors: how falsehood spread, why we believe them, what can be done*. Farrar, Strauss and Giroux, New York, NY, 2009.
51. Taber, C.S. and Lodge, M. Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*, 3 (2006), 755-769.
52. van Koningsbruggen, G.M., Das, E. and Roskos-Ewoldsen, D.R. How self-affirmation reduces defensive processing of threatening health information: Evidence at the implicit level. *Health Psychology*, *28*, 5 (2009), 563-568.
53. Wemple, E. The Maddow-Politifact clash *The Washington Post*, Washington, DC, 2012.
54. Zaller, J.R. *The nature and origins of mass opinion*. Cambridge University Press, New York, 1992.