Misinformation and Its Correction

Chloe Wittenberg and Adam J. Berinsky

Fake news is big news. From the diffusion of rumors and conspiracies in the United States to the spread of disinformation by Russian troll farms, misinformation is a hot topic among academics and journalists alike. How can we understand and correct such misinformation? A logical starting point is to fight fiction with fact. Indeed, many proposed solutions to the problem of misinformation assume that the proper remedy is merely to provide more information. In this view, if citizens were only better informed, misinformation would lose its power. However, ample research suggests that the answer is not so simple. Misinformation may continue to endure post-correction for several reasons. First, corrections are rarely able to fully eliminate reliance on misinformation in later judgments. Even when people recall hearing a retraction, the original misinformation may still influence their attitudes and beliefs (what is known as the continued influence effect). Worse yet, people may come to believe in misinformation even more strongly post-correction. In particular, retractions that run counter to individuals’ prior attitudes may bolster beliefs in the original misinformation (what are known as worldview backfire effects). These worldview backfire effects have their roots in directionally motivated reasoning; individuals process misinformation and corrections through the lens of their preexisting beliefs and partisan attachments, so they may actively dispute corrections that contradict their broader worldviews.

Although political misinformation is not a new phenomenon, the topic has received renewed attention in recent years, in conjunction with sweeping changes in the contemporary media environment. As the Internet and, particularly, social media become an increasingly common source for political information (Shearer and Matsa 2018), citizens receive more and more of their news in an uncontrolled and minimally regulated setting where misinformation may easily spread (Vosoughi, Roy, and Aral 2018). Validating these concerns, numerous studies of “fake news” spotlight social media platforms, including both Facebook and Twitter, as the primary incubators of misinformation.
during the 2016 US presidential election (e.g., Allcott and Gentzkow 2017; Guess, Nyhan, and Reifler 2020). However, even if the sources of misinformation have fundamentally changed, best practices for correcting misinformation have not. While many of the pieces cited in this chapter do not focus explicitly on the Internet or social media, these works can still inform scholarly understanding of how to correct misinformation on these platforms. The cognitive processes we highlight are likely to translate to the digital realm and are thus crucial to understand when developing prescriptions for social media–based misinformation. Nevertheless, we also spotlight a number of recent studies that examine methods for correcting misinformation in the context of social media.

**PREVIOUS REVIEW PIECES**

Several excellent review articles have already greatly enriched our knowledge of misinformation and its correction. Each is a valuable resource for deeper reading on this subject. In the interest of not rehashing existing work, we have made a conscious choice to showcase topics not already covered in these reviews. However, for the benefit of the reader, we summarize the primary takeaways from each piece and preview how we build on the groundwork they laid. First, Lewandowsky et al. (2012) provide a comprehensive summary of the literature on misinformation and its correction. In particular, they delve into the psychological roots of the continued influence effect and backfire effects and recommend appropriate interventions for practitioners seeking to mitigate these effects. However, since the article’s publication in 2012, the field has evolved in notable ways—especially regarding the existence and magnitude of different types of backfire effects. Swire and Ecker (2018) thus provide an updated summary of the literature and offer several new strategies for effectively correcting misinformation. We pick up where these two articles leave off; we discuss newer research on both the continued influence effect and backfire effects and suggest ways for future work to continue to flesh out these topics in even greater detail.

Second, Flynn, Nyhan, and Reifler (2017) offer a more recent review of the misinformation literature, with a specific focus on the relationship between directionally motivated reasoning and political misperceptions. In their view, individuals’ preexisting beliefs strongly affect their responses to corrections, such that individuals with different partisan or ideological leanings may respond to the same political facts in profoundly different ways. Importantly, the authors document a number of individual and contextual moderators of directionally motivated reasoning that predispose certain subsets of the population to be more vulnerable to worldview backfire effects. However, motivated reasoning is not the sole reason why misinformation persists over time. As studies of the continued influence effect demonstrate, individuals may continue to hold misinformed beliefs post-correction, even in the absence of
strong prior attitudes. As such, we take a closer look at the full range of psychological mechanisms that may impede attempts to correct misinformation.

Finally, Tucker et al. (2018) focus on the interplay of social media and political polarization in enabling the spread of misinformation, with particular emphasis on the specific actors who manufacture misinformation. However, though they present a wide-ranging and thorough analysis of the contemporary research on the production of digital misinformation, these authors devote substantially less attention to best practices for correction. As a complement to this work, we discuss recent research on strategies to correct misinformation appearing on social media platforms, including Facebook and Twitter.

**Organization of the Chapter**

In this chapter, we synthesize recent work on misinformation and its correction. Knowledge of this subject is still rapidly developing, and many questions remain unanswered and unresolved. Here, we pay particular attention to one of these important questions: Why does misinformation persist even after it has been corrected? To this end, we first provide a definition of misinformation and specify the core criteria that help to discriminate between the many related concepts in this area. Second, we discuss two key perspectives on the perseverance of misinformation post-correction: backfire effects and the continued influence effect. Third, we outline a number of individual and contextual moderators that might make certain individuals or groups especially susceptible to misinformation. Finally, we conclude with a series of recommendations for future research.

**Defining Misinformation: Mapping Key Criteria**

To understand how best to tackle the problem of misinformation, it is essential to first define what this term means. However, scholarly notions of what constitutes misinformation often differ significantly across works and across disciplines. These definitions are highly variable, ranging from simple statements about the misleading nature of misinformation to commentaries on the motivation for the spread of misinformation. Some scholars broadly characterize misinformation as false information; for example, Fetzer (2004) defines it as “false, mistaken, or misleading information” (p. 231), and Berinsky (2017) defines it as “information that is factually unsubstantiated” (p. 242). Other scholars take a more restricted view, contrasting the term with other concepts, such as disinformation. For instance, Wardle (2018) argues that

---

1 Indeed, though we attempt to provide a comprehensive review of the literature on misinformation correction, the field is moving so fast that this review may soon be out of date.
misinformation is “information that is false, but not intended to cause harm” (p. 5), whereas disinformation is “false information that is deliberately created or disseminated with the express purpose to cause harm” (p. 4). Finally, a third approach emphasizes the temporal nature of misinformation processing, arguing that misinformation’s primary feature is that it is first presented as true but later revealed to be false. Ecker et al. (2015) state that misinformation is “information that is initially presented as factual but subsequently corrected” (p. 102). Similarly, Lewandowsky et al. (2012) define misinformation as “any piece of information that is initially processed as valid but that is subsequently retracted or corrected” (pp. 124–125). In this sense, information only becomes misinformation when it is first believed and later corrected, separating misinformation from other false information that goes unrebutted.

Compounding the problem is the fact that the term “misinformation” is often confounded with other similar concepts. For instance, as noted in the previous paragraph, some authors attempt to draw a line between misinformation and disinformation, or “information that is false and deliberately created to harm a person, social group, organization, or country” (Wardle and Derakhshan 2017, p. 20). Other scholars speak of misperceptions, or “cases in which people’s beliefs about factual matters are not supported by clear evidence and expert opinion” (Nyhan and Reifler 2010, p. 305). Still others make reference to conspiracy theories, which offer unconventional explanations of the causes of events in terms of the “significant causal agency of a relatively small group of persons – the conspirators – acting in secret” (Keeley 1999, p. 116; see also Oliver and Wood 2014). Similar to, though broader in scope than, conspiracy theories are political rumors, which are “unverified stories or information statements people share with one another” (Weeks and Garrett 2014, p. 402). Finally, since the 2016 US presidential election, there has been much talk of fake news, which shares many similarities with disinformation but differs in its presentation. In particular, recent work defines fake news as “fabricated information that mimics news media content in form but not in organizational process or intent” (Lazer et al. 2018, p. 1094).

The multitude of definitions of misinformation speaks to the need for clarity on what exactly we, as a scholarly community, mean when we talk about misinformation. In an attempt to provide such structure, we compiled a wide variety of definitions of misinformation and related terms. Looking for common threads, we identified four overarching criteria for differentiating types of misinformation. First, we found that different definitions of misinformation place more or less emphasis on the truth value of the information – that is,

---

2 We are certainly not the first to propose such a typology (see Born and Edgington 2017; Tucker et al. 2018; Wardle 2018). However, we take a more comprehensive view than many of these previous works in that we seek to integrate a larger number of related concepts into a common theoretical framework.
whether the information has been proven to be untrue or whether it is merely unsubstantiated. Second, we noted that definitions of misinformation vary in their area of focus, particularly whether they emphasize the effects of false information versus false beliefs. Third, we found that scholars distinguish forms of misinformation based on their format, including whether or not the presentation of the information is designed to resemble traditional news sources. Finally, we noted differences in the perceived intentions of the actors who spread misinformation, in terms of their level of awareness that the information was false.

FOUR KEY CRITERIA

First, truth value: All forms of misinformation, at least to some degree, rest on shaky factual foundations. That is, all misinformation is in some way inaccurate. In some cases, misinformation is characterized by a lack of conclusive evidence to support a particular position, whereas, in others, it involves statements that run counter to mainstream consensus or expert opinion. However, the extent to which information is untrue varies across forms of misinformation; some subtypes may be definitively false (e.g., disinformation or fake news), whereas others may be merely misleading or unverified (e.g., political rumors).

Second, the area of focus: It is important to separate the presence of false information (misinformation) from the endorsement of false beliefs (misperceptions). This distinction is valuable because, as Thorson (2015) highlights, misperceptions are not exclusively caused by misinformation. Even if individuals only encounter true information, they may still arrive at inaccurate beliefs for other reasons, such as cognitive biases or misinterpretation of available facts. In this sense, the appropriate tools for correction may depend heavily on whether false beliefs are the clear product of misinformation or if they instead originate via other channels.

Third, format: Different types of misinformation may be presented in different ways. In some cases, misinformation may be embedded within otherwise accurate reports, whereas, in other cases, it may exist as standalone content. This is especially relevant to the study of fake news, or fabricated articles that imitate the appearance of traditional news stories (Allcott and Gentzkow 2017; Lazer et al. 2018). Fake news is a form of disinformation, as it is spread despite being known to be false, but it may be distinguished from other types of disinformation by its unique format – namely, its emulation of legitimate media outlets (Pennycook and Rand 2018). In addition to fake news, recent work also looks beyond textual forms of misinformation to other types of media, including manipulated images and videos (Kasra, Shen, and O’Brien 2016; Schwarz, Newman, and Leach 2016; Shen et al. 2019).

Finally, intentionality: Does the person transmitting misinformation sincerely believe it to be true or are they aware that it is false? By most
accounts, this is the primary means of distinguishing between misinformation and disinformation (for a review, see Wardle 2018). On the one hand, misinformation may circulate without any intent to deceive. For instance, in the wake of breaking news events, people increasingly turn to the Internet, and especially social media, for real-time updates. As new information is released in a piecemeal fashion, individuals may inadvertently propagate information that later turns out to be false (Nyhan and Reifler 2015a; Zubiaga et al. 2016). On the other hand, disinformation is false or inaccurate information that is deliberately distributed despite its inaccuracy (Stahl 2006; Born and Edgington 2017). People may choose to share fictitious stories, even when they recognize that these stories are untrue. Why might people knowingly promulgate false information? One answer relates to the disseminators’ motivations; although misinformation is typically not designed to advance a particular agenda, disinformation is often spread in service of concrete goals. For instance, fake news is often designed to go viral on social media (Pennycook and Rand 2018; Tandoc, Lim, and Ling 2018), enabling rapid transmission of highly partisan content and offering a reliable stream of advertising revenue (Tucker et al. 2018). In practice, however, determining a person or group’s intentions is extremely difficult. It is hard to uncover people’s “ground truth” beliefs about the veracity of a piece of information, and it is even harder to ascertain their underlying motivations. That said, recognizing the range of motivations for spreading misinformation is valuable, even if these motivations are hard to disentangle in the wild.

For the purposes of this chapter, we consider “misinformation” an umbrella term under which many associated concepts are subsumed. Moving forward, we recommend misinformation as the default term to use, unless explicitly referring to one of these more specific constructs. Given the difficulty of proving the motivations underlying the spread of false information, we adopt an intent-agnostic approach; we make no assumptions about what compels individuals or groups to broadcast misinformation. Instead, we take the view that misinformation – in all of its forms – may have a considerable, harmful impact on people’s beliefs and behavior. As such, in the discussion that follows, we cite examples of corrective strategies targeted at all different types of misinformation.

**Responses to Corrections: Continued Influence and Backfire Effects**

Detailing types of information is not a mere technical exercise. A well-functioning democratic society does not necessarily need to be guided by fully informed citizens, but an environment rife with misinformation can easily derail democracy. An uninformed citizenry is arguably far less pernicious than a misinformed citizenry (Kuklinski et al. 2000); as Hochschild and Einstein
people’s unwillingness or inability to use relevant facts in their political choices may be frustrating, but people’s willingness to use mistaken factual claims in their voting and public engagement is actually dangerous to a democratic polity” (p. 14). When the public holds misinformed beliefs, this can not only affect their individual attitudes and behaviors but also shape large-scale policy outcomes (e.g., health care reform, see Nyhan 2010; Berinsky 2017). Correcting misinformation is therefore a worthy goal; but how can it best be accomplished?

Previous research suggests that not all corrections are effective in reducing individuals’ reliance on misinformation. There are two pathways through which misinformation might continue to shape attitudes and behaviors post-correction: the continued influence effect and backfire effects. Engrained in the former is the notion that corrections are somewhat, but not entirely, effective at dispelling misinformation. More concerning, however, are the latter, in which corrections not only fail to reduce but actually strengthen beliefs in the original misinformation. Neither of these phenomena offers a particularly sanguine take on the ability to curtail the spread of misinformation. However, each offers its own unique predictions about the most promising avenues for corrections. We begin by reviewing the extant literature on backfire effects and then turn to the continued influence effect.

**Backfire Effects**

Providing factual corrections of misinformation may, under certain circumstances, only make things worse. Specifically, retractions that challenge people’s worldviews may entrench beliefs in the original misinformation. This phenomenon is known as a backfire effect or, more precisely, a worldview backfire effect. Nyhan and Reifler (2010) sounded the first alarm bells about the possibility of these worldview backfire effects. Across a series of studies, they found that, when certain subjects were presented with factual corrections that contradicted their political beliefs, they responded by becoming more, rather than less, wedded to their previous misperceptions. Since their highly influential piece was published, concerns about worldview backfire effects have taken hold both in popular media and in academic circles. This widespread interest has spawned an entire line of work dedicated to elucidating the psychological mechanisms that drive these effects.

Worldview backfire effects can be understood as a product of directionally motivated reasoning (for a comprehensive review, see Flynn et al. 2017). According to theories of motivated reasoning, individuals are motivated to process information in ways that align with their ultimate goals (Kunda 1990). In particular, individuals must balance several competing impulses,

---

3 Other terms for worldview backfire effects include “boomerang effects” (Hart and Nisbet 2012; Garrett, Nisbet, and Lynch 2013; Zhou 2016) or “backlash” (Guess and Coppock 2018).
including directional goals (to attain a desired outcome) and accuracy goals (to reach the correct conclusion). Worldview backfire effects transpire when directional motivations take precedence over accuracy goals – a frequent occurrence in the realm of politics (Lodge and Taber 2013).

Two complementary processes are at the heart of these effects. First, confirmation bias: Individuals tend to seek out and interpret new information in ways that validate their preexisting views. Along these lines, individuals also tend to perceive congenial information as more credible or persuasive than opposing evidence (Guess and Coppock 2018; Khanna and Sood 2018). Second, disconfirmation bias: When exposed to ideologically dissonant information, individuals will call to mind opposing arguments (counterarguing). In combination, these two processes can cultivate worldview backfire effects; when individuals are confronted with a correction that contradicts their past beliefs, they will act to both discount the correction and bolster their prior views.

Several studies have investigated the potential for worldview backfire effects in the context of misinformation. Although Nyhan and Reifler issued the earliest warnings about this phenomenon, it has since been reproduced across other settings. First, worldview backfire effects have been tied to message presentation, with individuals most resistant to message framing that contradicts their broader worldviews (Zhou 2016). Second, worldview backfire effects have been linked to source cues. For instance, several studies find that Republicans are averse to corrections from Democratic elites (Berinsky 2017) or nonpartisan fact-checking sources (Holman and Lay 2019). Finally, worldview backfire effects extend to the behavioral realm; across multiple studies, exposure to pro-vaccine corrections decreased future vaccination intentions among those already hesitant to get vaccinated (Skurnik, Yoon, and Schwarz 2007; Nyhan et al. 2014; Nyhan and Reifler 2015b; but see Haglin 2017).

Empirical studies have also taught us about the mechanisms that undergird worldview backfire effects. Consistent with a motivated reasoning perspective, worldview backfire effects appear rooted in counterarguing. In one experiment, Schaffner and Roche (2017) examine differences in survey response times following the release of the October 2012 jobs report, which announced a sharp decrease in the unemployment rate under the Obama administration. They find that those Republicans who took longer to provide estimates of the unemployment rate after the report’s release were less accurate in their responses, suggesting that worldview backfire effects may arise out of deliberate, effortful processes. However, more work beyond this initial study is certainly needed to isolate the mechanisms that underlie worldview backfire effects.

Counterarguing typically involves generating arguments to dispute a correction. Inverting this process, Chan et al. (2017) also find that corrections are generally less effective when people are asked to record arguments in favor of the original misinformation.
In light of mounting concerns about the potential for worldview backfire effects, scholars have explored several tactics for correcting misinformation while circumventing these effects. Although many routes to correction are possible, all designed to counteract directionally motivated reasoning, we summarize here two main subcategories of these corrections focused on source credibility and worldview affirmation.

First, the source of misinformation – as well as its correction – may have a profound impact on responses to corrections. When evaluating the accuracy of a claim, individuals rely heavily on source cues (Schwarz et al. 2016), which signal a source’s expertise or trustworthiness. In terms of expertise, if sources are depicted as authorities on a given subject, they are likely to be deemed more credible (Vraga and Bode 2017). In fact, expert consensus is considered a key “gateway belief” that can override directional impulses. For example, communicating the broad scientific agreement about climate change reduces partisan differences in climate change attitudes (van der Linden et al. 2015; van der Linden et al. 2017; Druckman and McGrath 2019; but see Kahan, Jenkins-Smith, and Braman 2011). However, the trustworthiness of a source seems to matter even more than expertise when countering misinformation (McGinnies and Ward 1980; Guillo and Geraci 2013). People are more likely to view sources as trustworthy if they share similar traits. As a result, corrections that are attributed to an in-group member (e.g., a leader of one’s preferred party) may be more effective than those credited to an out-group member (e.g., an opposing partisan, see Swire, Berinsky et al. 2017). Furthermore, though corrections are most frequently issued by elites, individuals are also receptive to corrections from members of their social circles (Margolin, Hannak, and Weber 2018; Vraga and Bode 2018), who may not be experts but may still be deemed trustworthy. Finally, trustworthiness is a function of one’s perceived stake in an issue. Recent research on “unlikely sources” (Berinsky 2017; Benegal and Scruggs 2018; Wintersieck, Fridkin, and Kenney 2018; Holman and Lay 2019) indicates that corrections are most persuasive when they come from sources who stand to benefit from the spread of misinformation (e.g., a Democratic politician or left-leaning publication correcting a fellow Democrat).

Second, where possible, corrections should be tailored to their target audience: the subset of people for whom these corrections would feel most threatening. Framing corrections to be consonant with, rather than antagonistic to, this group’s values and worldviews may thus be a successful corrective strategy (Kahan et al. 2010; Feinberg and Willer 2015; for reviews, see Lewandowsky et al. 2012; Swire and Ecker 2018). In a similar vein, some scholars propose using self-affirmation exercises (Cohen, Aronson, and Steele 2000; Cohen et al. 2007) to subdue directional motivations (Trevors et al. 2016; Carnahan et al. 2018); if people feel validated in their global self-worth,
corrections that impugn their political views may provoke a less defensive response (but see Nyhan and Reifler 2019).

BACKLASH AGAINST WORLDVIEW BACKFIRE EFFECTS

The previous discussion presumes that worldview backfire effects are not only dangerous but prevalent. Recently, however, there has been backlash against the very notion of worldview backfire effects, with some suggesting they are extremely rare in practice. Wood and Porter (2019) attempt to detect worldview backfire effects across a wide array of divisive issues and find little evidence to support their existence, even when using language identical to Nyhan and Reifler (2010). When examining several highly polarized issues, including gun control and capital punishment, Guess and Coppock (2018) likewise fail to uncover any worldview backfire effects in response to counter-attitudinal information. Instead, individuals seem to accommodate novel information into their later assessments of issues, even if that information runs counter to their beliefs (see also Porter, Wood, and Kirby 2018). However, even if corrections largely improve belief accuracy, these messages seem to have little impact on individuals’ subsequent attitudes, evaluations of politicians, or policy preferences (Swire, Berinsky et al. 2017; Aird et al. 2018; Nyhan et al. 2019; Porter, Wood, and Bahador 2019; Barrera et al. 2020).

What accounts for the discrepancies in results across studies? The answer may be both theoretical and methodological. First, there is a large body of work on the theoretic side. Some scholars have suggested that worldview backfire effects are more likely when corrections necessitate attitude change versus only pertain to a single, specific event (Ecker et al. 2014; Ecker and Ang 2019). Why are people better able to absorb ideologically dissonant corrections for one-off events? To answer this question, Ecker and Ang (2019) draw on stereotype subtyping theory (Richards and Hewstone 2001). According to this theory, individuals possess stereotypes about the customary behavior of different groups (e.g., members of a political party). Subtyping is a common response when group members act in ways that flout a well-established stereotype; rather than forming a new stereotype, individuals instead label contrary cases as exceptions to the broader rule. If misinformation pertains to a single, isolated event, individuals may thus be able to internalize disconfirming corrections without altering their deep-seated worldviews. It is much harder, however, to dismiss general patterns of behavior as anomalous. As such, people are more likely to resist corrections that, on acceptance, would require a large-scale shift in their core beliefs.

Redlawsk, Civettini, and Emmerson (2010) provide a somewhat different perspective on the boundaries of worldview backfire effects. They posit an “affective tipping point” at which individuals cease to engage in motivated reasoning and instead revise their beliefs to be more accurate – in other words, the point at which individuals pivot from directional to accuracy
goals. As people encounter more and more disconfirming information, they may reach a critical threshold at which they are no longer motivated to defend their previous views. In this view, worldview backfire effects will occur until enough contradictory evidence accumulates. After this point, individuals will begin to rationally update their beliefs in response to corrections rather than double down on their previously misinformed views. This theoretical account generates somewhat divergent predictions from Ecker and Ang. For both sets of authors, general cases of misinformation should provoke heightened discomfort. However, if enough contrary evidence comes to light, Redlawsk and colleagues anticipate a diminished likelihood of worldview backfire effects, whereas Ecker and Ang seem to predict the exact opposite. Further work may be needed to adjudicate between these two explanations. In particular, efforts to pinpoint the precise location of this tipping point may prove fruitful.

Methodological differences may play a role as well. Worldview backfire effects may not be immediately apparent post-correction. Instead, they may only emerge after some time has elapsed (Peter and Koch 2016; Pluviano, Watt, and Della Sala 2017). However, most research on worldview backfire effects just measures the effect of corrections after a short distraction task. In contrast, studies that do incorporate lengthy time delays (e.g., Berinsky 2017; Swire, Berinsky et al. 2017) find that the benefits of corrections quickly dissipate. Worldview backfire effects may therefore only be visible after a delay. Studies of these effects should therefore aim to measure responses at multiple points in time. In addition, worldview backfire effects are more probable for high-salience issues where individuals have strong prior attitudes (Flynn et al. 2017). Nevertheless, the deep-rooted nature of these issues may limit the range of effect sizes that a single experimental manipulation can elicit. As a result, worldview backfire effects may be especially hard to detect for highly polarized issues – the very issues where we would expect the most pervasive effects.

On a related note, it is essential to come to some consensus regarding what, exactly, we consider a “backfire effect.” In particular, worldview backfire effects may be an artifact of the baseline against which they are measured. Scholars generally define worldview backfire effects as cases where the presentation of both misinformation and its correction is worse than presenting misinformation uncorrected. Yet when considering the deleterious effects of misinformation in society, a more expansive definition may be appropriate. Worldview backfire effects are commonly measured experimentally by comparing respondents who were exposed to misinformation and corrections. However, when examining information that has already spread through society – beliefs about President Obama’s citizenship, for example – a better baseline might be people’s beliefs if they had not been reexposed to misinformation as part of an experiment. If providing a correction to misinformation is worse than providing no information at all, strategies for mitigating misinformation may require substantial adjustment.
CONTINUED INFLUENCE EFFECT

The ubiquity of worldview backfire effects remains an open question. However, even if these effects are overblown, valid concerns about the unintended consequences of corrections remain. In particular, the format in which corrections are delivered may bolster beliefs in misinformation, even in the absence of worldview backfire effects. A near-universal finding in the misinformation literature is that, even after its correction, misinformation continues to influence people’s attitudes and beliefs (for a review, see Walter and Tukachinsky 2019). This is known as the continued influence effect (Wilkes and Leatherbarrow 1988; Johnson and Seifert 1994). Importantly, people may correctly recall a retraction yet still use outdated misinformation when reasoning about an event. From this perspective, corrections can partially reduce misperceptions but cannot fully eliminate reliance on misinformation in later judgments.

Why does misinformation linger post-correction? Scholars suggest two potential reasons for the continued influence effect. First, according to the mental model theory, individuals construct models of external events in their heads, which they continuously update as new information becomes available (Johnson and Seifert 1994; Swire and Ecker 2018). However, retractions often threaten the internal coherence of these models (Gordon et al. 2017; Swire and Ecker 2018). As a result, even if individuals explicitly recall corrections, they may nevertheless continue to invoke misinformation until a plausible alternative takes its place. Numerous studies find that corrections are more effective when they contain alternative causal accounts rather than just negate the original misinformation (Johnson and Seifert 1994; Ecker, Lewandowsky, and Tang 2010; Nyhan and Reifler 2011; but see Ecker et al. 2015).

Secondly, the continued influence effect can be understood through dual-process theory. Dual-process theory distinguishes between two types of memory retrieval: automatic and strategic. Automatic processing is fast and unconscious, whereas strategic processing is deliberate and effortful. In addition, automatic processing is relatively acontextual, distilling information down only to its most essential properties, whereas strategic processing is required to retrieve specific details about a piece of information (Ecker et al. 2011). As a result, individuals may be able to remember a piece of misinformation but not recall relevant features, such as its source or perceived accuracy (Swire and Ecker 2018). In this view, the continued influence effect constitutes a form of retrieval failure; misinformation is automatically retrieved, but its retraction is not. This emphasis on automatic versus strategic processing is also consistent with an online processing model of misinformation (Lodge and Taber 2013; Thorson 2016). According to this model, initial misinformation is encoded with a stronger affective charge than its correction, meaning that misinformation
will continue to dominate subsequent evaluations until individuals engage in the strategic processing necessary to explicitly recall a correction.

Most direct studies of the continued influence effect use variants of the same research design, based on the “warehouse fire” script (Wilkes and Leatherbarrow 1988; Johnson and Seifert 1994). In this scenario, the cause of a fire is initially attributed to volatile chemicals stored in a closet, but the closet is later revealed to have been empty. Subsequent studies have adapted this narrative to other contexts, such as police reports or political misconduct, but all follow a similar format in which information about a breaking news event is relayed over a series of short messages. Experimenters randomly assign some subjects to read a critical piece of misinformation (e.g., the presence of flammable materials) as well as its retraction (e.g., the empty closet). However, this communication technique is arguably ill-suited to the study of political misinformation. First, many of these studies present misinformation and corrections as coming from the same source. However, in the realm of politics, the sources most likely to issue corrections may be the ones least likely to spread the misinformation in the first place. Second, the sequencing of messages may not accurately mimic how individuals encounter information in the real world, where the temporal distance may be either much shorter (instantaneous, if people see a correction before or concurrently with the original misinformation) or much longer (if corrections are issued at a later date). Finally, these studies usually rely on fictional scenarios that do not implicate social identities or prior attitudes (but see Ecker et al. 2014), both of which may increase the likelihood of the continued influence effect.

Accordingly, recent work has sought to investigate the presence of the continued influence effect in the political domain. Most notably, Thorson (2016) introduces the concept of “belief echoes,” a version of the continued influence effect focused on attitudes rather than causal inferences. According to her theory, misinformation may continue to influence political attitudes through two separate processes. First, automatic belief echoes develop as a byproduct of online processing. Even when individuals accept corrections as true, misinformation may still be automatically activated, thereby continuing to affect attitudes outside of conscious awareness. Deliberative belief echoes, on the other hand, occur when individuals assume that the existence of one piece of negative information—even if it is known to be false—increases the likelihood that other relevant negative information is true (a “where there’s smoke, there’s fire” philosophy). Together, these automatic and deliberative belief echoes may contribute to the perpetuation of misinformation post-correction.

FAMILIARITY BACKFIRE EFFECTS

The continued influence effect suggests that corrections are somewhat, though not entirely, effective in reducing belief in misinformation. In fact, contrary to worldview backfire effects, the continued influence effect does not require the
existence of strong prior attitudes. However, backfire effects might occur even in the absence of worldview threat. In particular, corrections that repeat misinformation may amplify its influence, constituting an alternate form of backlash known as familiarity backfire effects. Of note, within the political science literature, the term “backfire effect” almost exclusively refers to worldview backfire effects. However, familiarity backfire effects are a much more common area of focus within the psychology literature. To avoid confusion, we treat these concepts as separate phenomena.

Familiarity backfire effects involve cases in which retractions increase, rather than reduce, reliance on misinformation by making misinformation feel more familiar. These effects are primarily studied in the context of repetition. In particular, familiarity backfire effects are considered the product of the illusory truth effect, wherein “repeated statements are easier to process, and subsequently perceived to be more truthful, than new statements” (Fazio et al. 2015, p. 993). The illusory truth effect operates through a series of complementary psychological mechanisms. First, repeating information strengthens its encoding in memory, enabling easier retrieval later on (for reviews, see Lewandowsky et al. 2012; Peter and Koch 2016). Second, the difficulty with which information is processed influences its perceived authenticity. This is tied to the metacognitive experience of “processing fluency” (Schwarz, et al. 2007); information that is easier to process feels more familiar, and familiarity is a key criterion by which individuals judge accuracy (Alter and Oppenheimer 2009). Accordingly, if individuals have repeated contact with a piece of misinformation, they may perceive it as more credible than if they encounter it only once, regardless of its content.

The illusory truth effect is of particular concern in regard to misinformation correction, given the standard format of corrections. In particular, as part of the debunking process, most corrections directly reference the original misinformation. For instance, the commonly employed “myths vs. facts” strategy involves repeating misinformation (the “myth”) while simultaneously discrediting it (the “fact”). As such, repeated exposure to misinformation – even during its correction – may activate the familiarity heuristic and therefore enhance the perceived accuracy of misinformation. Indeed, familiarity backfire effects have been detected across numerous studies of this specific correction style (Schwarz et al. 2007; Peter and Koch 2016; but see Cameron et al. 2013).

Familiarity backfire effects may be especially prominent after a time delay. Though individuals are typically able to differentiate fact from fiction immediately after viewing a correction, they may soon forget the details of the correction and retain only the gist of the original misinformation. For example, Skurnik et al. (2007) find that subjects were able to distinguish between myths and facts about the flu vaccine right after reading an informational flyer but, after only a short break, were significantly more likely to mistake myths for facts than the reverse. In addition, in a study of healthcare reform, Berinsky (2017)
notes that the effectiveness of corrections faded rapidly over time, with subjects exposed to corrections no more likely than those in a control group to reject a rumor about “death panels” after just a week. Even if corrections are initially able to reduce misperceptions, their benefits may be short-lived.

Familiarity backfire effects are also likely to be relatively universal, as the illusory truth effect is largely robust across individuals and situations. Even when they have prior knowledge about a subject, individuals tend to rate repeated statements as truer than new statements (Fazio et al. 2015). In addition, the illusory truth effect is only modestly associated with dispositional skepticism (DiFonzo et al. 2016) and is uncorrelated with several psychological traits, such as analytical thinking and need for closure, that are otherwise connected to the processing of misinformation (De keersmaecker et al. 2020). Finally, the illusory truth effect appears independent of motivated reasoning; across both politically consistent and discordant statements, repeated exposure corresponds to higher accuracy ratings (Pennycook, Cannon, and Rand 2018).

Not all scholars, though, have found evidence of familiarity backfire effects. Although most scholars acknowledge that familiarity affects the processing of corrections, some dispute the negative relationship between repetition of misinformation and belief accuracy (e.g., Swire, Ecker, and Lewandowsky 2017; Pennycook et al. 2018). In fact, Ecker, Hogan, and Lewandowsky (2017) find that retractions that include reminders of the original misinformation are more effective than retractions without this repetition. They attribute these results to the benefits of coactivating misinformation and corrections (see also Swire and Ecker 2018). When misinformation and its correction are summoned simultaneously, individuals are better able to detect discrepancies between the original misinformation and the factual evidence. This “conflict detection” expedites the knowledge revision process, leading to more efficient belief updating. In light of these contradictory findings, it remains unclear how concerned we should be about familiarity backfire effects when correcting misinformation. However, we discuss a number of strategies in the following section to minimize the risk of these effects, regardless of their prevalence.

AVOIDING FAMILIARITY BACKFIRE EFFECTS

What strategies exist to correct misinformation while evading familiarity backfire effects? The most obvious solution is to focus on the correction
without alluding to the original misinformation. However, several of the pieces cited in this chapter suggest that avoiding repetition is not a magic bullet; at times, providing details about a piece of misinformation can aid in the correction process. Moreover, even if avoiding repetition is the goal, this may not always be possible. In many cases, misinformation is published by one source and corrected by another. Rather than just affirm the facts, corrections may need to invoke the original misinformation in order to provide proper contextualization. Instead of avoiding repetition of misinformation, it may thus be more valuable to focus on reiterating corrections, as a means of increasing the familiarity of accurate information (Ecker et al. 2011).

Furthermore, processing fluency is not solely a function of repetition (Schwarz et al. 2007). On the whole, information that is easier to process will be perceived as more familiar (and therefore more valid). Consequently, corrections may be more successful when they are less cognitively taxing. For example, visual corrections may be easier to digest than long-form fact-checking articles (Alter and Oppenheimer 2009; Schwarz et al. 2016). Previous studies using photographs (Garrett, Nisbet, and Lynch 2013), infographics (Nyhan and Reifler 2019), and videos (Young et al. 2018) largely corroborate this hypothesis (but see Nyhan et al. 2014). Similarly, corrections that employ simple words or grammatical structures may be more decipherable than linguistically complex corrections (Alter and Oppenheimer 2009). The readability of corrections is thus another important consideration for future research to explore. Finally, it may be optimal for corrections to combine multiple approaches. For instance, corrections that pair pithy images (e.g., PolitiFact’s Truth-O-Meter) with accompanying descriptive text may be especially effective (Amazeen et al. 2016).

VICTIMS OF MISINFORMATION: MODERATORS OF MISINFORMATION AND ITS CORRECTION

Overall, misinformation appears both pervasive and difficult to correct once it spreads. However, not all misinformation is created equal, nor are all individuals equally susceptible to its influence. Thus, it is important to examine which groups are most likely to be affected by misinformation in society. In the sections that follow, we outline several factors – both individual and contextual – that may affect the persistence of misinformation among certain groups, by making individuals either more likely to believe misinformation or more resistant to its correction.

INDIVIDUAL FACTORS

We first discuss several individual-level moderators of receptiveness to misinformation and responsiveness to corrections. These factors may be
bifurcated into two strands: those that are explicitly political in nature and those that reflect more fundamental personal or psychological orientations.

Political Factors

Two main political factors contribute to the nature and severity of misinformation effects: political sophistication and ideology. One of the most frequently studied moderators of correction effectiveness is political sophistication, which includes aspects of political knowledge, engagement, and education (for a review, see Flynn et al. 2017). At first glance, more politically sophisticated individuals should be less susceptible to misinformation than less-informed citizens, as they can draw on their superior knowledge to discern fact from fiction. Along these lines, Berinsky (2012) finds that more politically engaged individuals are, on the whole, more likely than others to reject political rumors. However, he also finds that politically sophisticated Republicans are more likely to accept rumors about Democrats, suggesting that political knowledge does not entirely inoculate individuals against misinformation.

In fact, belief in misinformation may actually be more prevalent within this more educated and engaged group. Recent research finds that individuals who are more politically active and engaged are more likely to share misinformation via social media, thereby contributing to the spread of misinformation to other members of the public (Valenzuela et al. 2019). Moreover, politically sophisticated individuals may be more resistant to corrections. In general, politically sophisticated individuals tend to evince the strongest directional motivations (Lodge and Taber 2013), corresponding to greater endorsement of misinformation that reinforces their prior beliefs (Nyhan, Reifler, and Ubel 2013; Miller, Saunders, and Farhart 2016; Jardina and Traugott 2019). Nyhan and Reifler (2010) propose two mechanisms by which this might be the case: a biased information search and biased information processing. First, politically sophisticated individuals may be more likely to selectively consume ideologically consistent media (confirmation bias), thereby filtering out the sources most likely to publish attitude-incongruent corrections. Second, when encountering attitude-incongruent corrections, politically sophisticated individuals may be best equipped to counterargue against these corrections (disconfirmation bias).

The most politically sophisticated individuals seem the least amenable to corrections when misinformation supports their preexisting beliefs. As a result, corrections may fail to reduce and may even enhance belief in misinformation among this small but consequential group. From this perspective, political sophistication is a crucial determinant of responses to misinformation and its correction. Highly sophisticated partisans have both the motivation and the expertise to discount corrections that run counter to their predispositions. Furthermore, less engaged citizens are unlikely to be exposed to corrections in
the first place. When considering solutions to the spread of misinformation, the standard prescription is merely to provide more information. However, this heightened susceptibility to misinformation among the most informed citizens exposes the limits to this approach; when individuals are knowledgeable about and involved in politics, this engagement may ironically engender the strongest opposition to corrections. Thus, a more informed populace may not be a panacea if corrections continue to heighten directional motivations.

**Political Ideology and Partisanship**

An active debate in the misinformation literature concerns potential asymmetries in responses to misinformation based on political ideology and partisan identification (for a review, see Swire, Berinsky et al. 2017). Specifically, some scholars claim that conservatives and Republicans are especially vulnerable to misinformation. In a widely cited article, Jost et al. (2003) catalog a laundry list of predictors of conservatism (e.g., close-mindedness, intolerance of ambiguity), many of which could engender openness to misinformation and resistance to corrections. In a later piece, Jost et al. (2018) highlight several other factors associated with conservatism, including an emphasis on in-group consensus and homogeneous social networks, that may give rise to “echo chambers” in which misinformation can easily spread (see also Nam, Jost, and Van Bavel 2013; Ecker and Ang 2019). Taken together, these pieces paint a picture of conservatives as resistant to change, averse to uncertainty, and drawn to one-sided information environments – all of which might predispose those on the right to favor misinformation, relative to their moderate or liberal counterparts.

These theoretical expectations have some empirical backing. Recent research finds that, during the 2016 election, Republicans were more likely than Democrats to read and share fake news (Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Guess et al. 2020). Furthermore, ideology and partisanship are associated with differences in responses to corrections. For example, Nyhan and Reifler (2010) report evidence of ideological asymmetry in responses to corrections. Although they find that, regardless of partisan leaning, corrections were generally less effective when they were attitude-incongruent, worldview backfire effects were visible for Republicans but not Democrats (see also Ecker and Ang 2019).6 These individual-level differences may be exacerbated by system-wide differences in conservative versus liberal media. Although misinformation originates in both liberal and conservative circles, the insular nature of the conservative media ecosystem may be more conducive to the spread of misinformation (Faris et al. 2017; see also Barberá, Chapter 3, this

---

6 This observed asymmetry, however, cannot be definitively ascribed to individual-level differences across ideological groups. For instance, there may be qualitative differences between conservative- and liberal-leaning misinformation that make the former stickier. In particular, Nyhan and
volume), and conservative media sources are more likely than liberal sites to dismiss or otherwise derogate nonpartisan fact-checkers (Iannucci and Adair 2017). Finally, these system-wide differences also extend to individual behavior. In an analysis of tweets about the 2012 presidential election, Shin and Thorson (2017) find that Republicans retweeted or replied much less frequently to fact-checking sites than Democrats—and their replies tended to be more acrimonious. Similarly, across both Facebook and Twitter, Amazeen, Vargo, and Hopp (2018) find that liberal-leaning individuals tend to be more likely than others to share fact-checking information.

However, this emphasis on the psychological profiles of political conservatives is not without controversy. Kahan and colleagues contend that motivated reasoning is not a uniquely right-wing phenomenon. Instead, all individuals are motivated to express and maintain beliefs similar to those of other members of their identity groups (the “cultural cognition thesis,” e.g., Kahan et al. 2011; Kahan 2013). In line with this perspective, several recent works suggest that liberals are not, in fact, immune to the effects of misinformation (Aird et al. 2018; Guess et al. 2019). Across numerous fields, ranging from science to politics, both conservatives and liberals evince similar levels of motivated reasoning (Nisbet, Cooper, and Garrett 2015; Meirick and Bessarabova 2016; Frimer, Skitka, and Motyl 2017; Swire, Ecker et al. 2017; Ditto et al. 2019). While conservatives may disproportionately display the motivational tendencies associated with belief in misinformation, these proclivities do not necessarily translate to behavioral differences.

While political knowledge has been firmly established as a key moderator of misinformation effects, via its relationship to directionally motivated reasoning, the jury is still out regarding the role of political ideology and partisanship. Although conservatives and Republicans may, under certain conditions, be more sensitive to misinformation than others, this divide may be overstated. Are observed cases of ideological asymmetry a function of deeply rooted psychological traits, or do they instead reflect systematic differences in conservative versus liberal media environments (or in the misinformation itself)? Future work should continue to grapple with this tricky distinction.

Personal and Psychological Factors

Misinformation, however, is not contained to the political sphere. More basic personal and psychological factors may predispose certain individuals to champion misinformation and disavow corrections across domains. We

Reifler’s (2010) experiments rely on actual examples of misinformation (stem cell research and weapons of mass destruction in Iraq). While this approach has the benefit of greater external validity, these cases may diverge in notable ways beyond their ideological slant (e.g., issue salience or importance). Studies that focus on fabricated misinformation, rather than real-world rumors, may thus be better suited to identifying potential partisan or ideological asymmetries.
highlight four of these potential moderators, namely age, analytical thinking, need for closure, and psychological reactance. Scholars highlight age as a key demographic variable influencing both exposure and responses to misinformation. Several recent studies find that older adults are more likely than others to share fake news stories on social media (Grinberg et al. 2019; Guess et al. 2019). However, other work finds that old age is also associated with greater sharing of fact-checks on social media (Amazeen et al. 2018), suggesting that older cohorts may engage differently with political content on social media, relative to their younger counterparts.

Scholars have also identified analytical thinking, or a person’s capacity to override gut feelings and intuitions, as another determinant of their responses to misinformation. In this sense, individuals who are more prone to careful, deliberate processing of information (or “cognitive reflection”) seem to be less susceptible to misinformation. Analytical thinking is associated with reduced belief in conspiracy theories (Swami et al. 2014) and increased accuracy in judging fake news headlines (Pennycook and Rand 2018; Bronstein et al. 2019; Pennycook and Rand 2020). Furthermore, highly analytical individuals are more willing than others to adjust their attitudes post-correction, even after controlling for a host of other variables (De keersmaecker and Roets 2017; see also Tappin, Pennycook, and Rand 2018). While most studies conceptualize analytical thinking as a dispositional trait, recent work suggests that interventions designed to encourage greater deliberation may also prove an effective tool for correcting misinformation (Bago, Rand, and Pennycook 2020).

Need for closure may also shape an individual’s susceptibility to misinformation. Need for closure refers to “the expedient desire for any firm belief on a given topic, as opposed to confusion and uncertainty” (Jost et al. 2003, p. 348, italics in original). This motivation fosters two main behavioral inclinations: the propensity to seize on readily available information and the tendency to cling to previous information (Jost et al. 2003; Meirick and Bessarabova 2016; De keersmaecker et al. 2020). Consequently, individuals with a high need for closure may be more trusting of initial misinformation, which provides closure through explaining the causes of events, and more resistant to corrections, which may sow feelings of confusion and uncertainty (Rapp and Salovich 2018). Need for closure, however, is primarily used as a control variable in studies of misinformation and is rarely the main construct of interest. Indeed, the few studies connecting a need for closure to misinformation focus solely on the endorsement, rather than correction, of misinformation (e.g., Leman and Cinnirella 2013; Moulding et al. 2016; Marchlewksa, Cichocka, and Kossowska 2018). Nevertheless, need for closure may also moderate the

---

7 Of course, these factors may be correlated with political sophistication and ideology and may therefore be at the root of some of the empirical regularities cited in the “Political Factors” section.
effectiveness of corrections. For instance, individuals with a high need for closure may be especially vulnerable to the continued influence effect; if these individuals are less acceptant of gaps in their mental models of an event, they may be more likely to retain misinformation in the absence of plausible alternative explanations. Moving forward, future research should continue to probe the extent to which a high need for closure predisposes certain individuals to disregard corrections.

Finally, high levels of psychological reactance may trigger backfire effects by stimulating counterarguing. Psychological reactance occurs when individuals perceive a threat to their intellectual or behavioral freedoms, such as when they feel strong pressure to adopt a certain attitude or belief (Sensenig and Brehm 1968). In short, many people do not like being told what or how to think. As a result, they may actively defy corrections that seem overly authoritative (Garrett et al. 2013; Weeks and Garrett 2014). Misperceptions may thus be even more difficult to remedy for individuals who eschew conformity. Indeed, across countries, anti-vaccination attitudes are significantly and positively correlated with psychological reactance (Hornsey, Harris, and Fielding 2018). Moreover, several studies document a link between psychological reactance and resistance to climate change messaging (Nisbet et al. 2015; Ma, Dixon, and Hmielowski 2019). A deeper focus on psychological reactance may therefore help reconcile previously perplexing findings in the misinformation literature. Some accounts of the continued influence effect posit that individuals continue to endorse misinformation because they do not believe corrections to be true (Guillory and Geraci 2013). This tendency may be heightened among those with a contrarian streak. In addition, several scholars caution against providing too many corrections (“overkill” backfire effects, see Cook and Lewandowsky 2011; Lewandowsky et al. 2012; Ecker et al. 2019). The purported perils of overcorrection may have their roots in psychological reactance (Shu and Carlson 2014); inundating people with a surfeit of corrections may provoke feelings of reactance, particularly among those already liable to reject consensus views.

**CONTEXTUAL FACTORS**

Along with individual-level moderators of misinformation effects, contextual factors may play an important role in guiding responses to misinformation and its correction. These variables include the **content** of misinformation as well as the **environments** in which misinformation is consumed and corrected.

**CONTENT-BASED FACTORS**

The actual substance of misinformation – including its subject matter and tone – is an important determinant of its correctability. First, corrections may
be differentially effective across issue areas. For example, in a meta-analysis of studies of misinformation correction, Walter and Murphy (2018) find that corrections are more effective for health-focused misinformation than for political and scientific misinformation. Second, misinformation may vary in its affective content. Negatively valued misinformation tends to be more durable than positive or neutral misinformation (Forgas, Laham, and Vargas 2005; Guillory and Geraci 2016; but see Mirandola and Toffalini 2016). Moreover, the emotions that misinformation arouses may also influence its persistence (Vosoughi et al. 2018). In particular, Weeks (2015) finds that feelings of anger tend to encourage directionally motivated processing of corrections, whereas feelings of anxiety tend to reduce partisan differences in responses to corrections. However, misinformation does not seem to inspire these emotions in equal measure. Text analysis of comments on Facebook posts containing misinformation finds that responses to misinformation are more frequently characterized by anger as opposed to anxiety (Barfar 2019).

ENVIRONMENTAL FACTORS

How people encounter misinformation may also influence both their contact with and their responses to corrections. Although misinformation is an age-old problem, the topic has garnered attention in recent years due to concerns about how the Internet – and especially social media – might extend its reach. Many producers of misinformation use social media sites as their main means of disseminating misinformation (Tucker et al. 2018). Reflecting this fact, several recent studies emphasize the role of social networking sites, including Facebook and Twitter, in amplifying exposure to fake news content (Allcott and Gentzkow 2017; Allcott, Gentzkow, and Yu 2019; Guess, Nyhan, and Reifler 2020). However, some work suggests that exposure to fake news on social media is limited to only a small subset of the population (Grinberg et al. 2019; Guess et al. 2019), and others find that social media use is only weakly associated with the endorsement of false information (Garrett 2019).

Even if misinformation may propagate easily via social media, these platforms may be essential to combating its spread. After all, social media can spread corrections in addition to misinformation (Vraga 2019). Much work focuses on efforts by social media sites to prevent the spread of misinformation or other harmful rhetoric in the first place (for reviews, see Guess and Lyons, Chapter 2, and Siegel, Chapter 4, this volume). However, social media platforms can also play an active role in correcting misinformation after the fact. To this end, scholars have studied the effectiveness of two types of social media–based corrections: algorithmic and social corrections. Some social media sites have built-in functionalities that can be deployed to combat misinformation. For example, Bode and Vraga (2015, 2018) focus on Facebook’s “related stories” feature, which recommends relevant articles underneath shared links, and find
that fact-checking articles publicized through this system may be effective in increasing belief accuracy – especially on issues where individuals do not possess strong prior attitudes. Another proposed form of algorithmic correction relies on “crowdsourced” data on the trustworthiness of different news outlets to decrease the likelihood that individuals will encounter posts from unreliable sources (Pennycook and Rand 2019). In addition to these algorithmic corrections, other social media users (e.g., Facebook friends or Twitter followers) can intervene to provide corrections (Vraga and Bode 2018). These social corrections may be especially effective, as individuals are more likely to accept corrections from people they already know (Friggeri et al. 2014; Margolin et al. 2018).

However, some scholars caution about the potential for social media to undermine the correction of misinformation. The “social” nature of social media may increase levels of exposure to misinformation, as individuals are more likely to read news that has been shared or endorsed by members of their social networks (Messing and Westwood 2014; Anspach 2017). The nature of the social media environment may also inhibit corrections of misinformation; Jun, Meng, and Johar (2017) warn that people are less likely to fact-check statements in social settings – a form of “virtual bystander effect.” Furthermore, even if corrections circulate on social media, individuals may be more attentive to user comments on these posts than to the actual fact-checking messages themselves. If these comments distort or otherwise misrepresent corrections, individuals may not become better informed, despite their exposure to fact-checking information (Anspach and Carlson 2018).

Finally, and most importantly, corrective efforts on social media may have unintended consequences. Given the difficulties of correcting misinformation postexposure, many scholars recommend preemptive interventions designed to induce skepticism prior to misinformation exposure (Ecker et al. 2010; Peter and Koch 2016; Cook, Lewandowsky, and Ecker 2017). Specifically, some scholars recommend training individuals to detect and resist misinformation by highlighting the techniques commonly deployed by creators of misinformation (Roozenbeek and van der Linden 2019a, 2019b). Social networking sites have adopted similar models. For example, after the 2016 US presidential election, Facebook rolled out a new system to flag potentially inaccurate stories as disputed or false. Warning labels of this sort may be effective in reducing the sharing of flagged stories (Mena 2019). However, false stories that go undetected by this system may be viewed as more accurate than they would have were the system never put in place (Pennycook et al. 2020). Similarly, general warnings about the potentially misleading nature of social media posts may decrease beliefs in the accuracy of true headlines (Clayton et al. 2019), suggesting that corrections issued on social media might inadvertently erode trust in credible media content.
EXPOSURE TO FACT CHECKS

Only a small subset of the population will likely encounter both misinformation and corrections. On the misinformation side, while some types of misinformation are widespread (e.g., the birther movement), many remain fringe beliefs. Despite rampant fears about “fake news,” fake news sites during the 2016 and 2018 elections received the bulk of their traffic from a very small set of highly partisan consumers (Grinberg et al. 2019; Guess et al. 2019, 2020). On the corrections side, a limited number of people view a limited number of corrections. Relatively few people ever visit professional fact-checking sites, such as PolitiFact or Factcheck.org, without external prompting; the public appreciates fact-checking in theory but shows little interest in practice (Nyhan and Reifler 2015c). These low levels of engagement are exacerbated by patterns of selective exposure to and sharing of fact-checking messages on social media platforms (Shin and Thorson 2017; Zollo et al. 2017; Hameleers and van der Meer 2020), as partisans tend to seek out and share fact checks that reinforce their prior attitudes. If highly engaged members of the public cherry-pick favorable fact-checking messages to share with others, those exposed to these messages may observe only a narrow, unrepresentative slice of the available set of corrections.

Finally, even if individuals do take the initiative to visit fact-checking sites, these sites frequently choose to cover markedly different topics. In fact, even when their coverage does overlap, fact-checking organizations often reach diametrically opposed conclusions about the factual basis for a given piece of information (Marietta, Barker, and Bowser 2015). These potential discrepancies are consequential, as several studies of fact-checking messages find that the content of these messages (e.g., affirming or refuting information) matters more than their source (e.g., Fox News, MSNBC, or PolitiFact) in increasing belief accuracy (Wintersieck 2017; Wintersieck et al. 2018).

CONCLUSION

Within both popular media and academia, concerns abound regarding the prevalence and persistence of misinformation. In an age where misinformation can diffuse rapidly via the Internet and social media, it is more imperative than ever to think creatively about how best to debunk misinformation. Although misinformation may take many forms—ranging from political rumors to disinformation—each of these forms presents a potential threat to democracy by distorting attitudes, behavior, and public policy. Definitional concerns should therefore take a backseat to mitigating the harmful effects of misinformation. Given the potential dangers of misinformation, devising effective strategies for correction is crucial, yet previous prescriptions have often come up short.
In this review, we have discussed two phenomena that may contribute to the durability of misinformation post-correction: the continued influence effect and backfire effects. Though scholars have found evidence that each of these processes undermines the effectiveness of corrections, recent works have cast doubt on their pervasiveness. In light of these findings, several areas merit further research. First, although worldview backfire effects may be less widespread than originally thought, the existence of these effects remains an open question. Efforts to isolate the conditions, both theoretical and methodological, under which worldview backfire effects are most likely to occur may help to resolve this ongoing debate. Similarly, though scholars frequently discourage the repetition of misinformation within corrections, more recent studies have cast doubt on the prevalence of familiarity backfire effects. Given that traditional methods of correction often cite the original misinformation, understanding whether and how this repetition might undercut their effectiveness is important. In particular, clarifying the conditions under which repetition is a benefit versus a hindrance may yield practical recommendations for improving the success of fact-checking sites. Finally, misinformation does not affect all individuals equally, nor is all misinformation equally persuasive. Continuing to identify these places of heterogeneity may enable more active targeting of corrections to those subgroups where misinformation is most likely to take root.

REFERENCES


Skurnik, I., Yoon, C., & Schwarz, N. (2007). “Myths & Facts” about the flu: Health education campaigns can reduce vaccination intentions. http://webuser.bus.umich.edu/yoon/research/Papers/Skurnik%5FYoon%5FSchwarz%5F2005%5FMyths%5FFacts%5FFlu%5FHealth%5FEducation%5FCampaigns%5FJAMA.pdf


