



When fools have greater influence than liars: testing the insidious bullshit hypothesis with the illusory truth effect

John V. Petrocelli¹ · Elijah N. Rice² · Joseph M. Curran¹

Received: 17 September 2024 / Accepted: 15 June 2026
© The Author(s) 2026

Abstract

In an era of viral misinformation and pervasive truth-indifferent communication, understanding what shapes everyday beliefs about what is true has never been more consequential. Little is known about the consequences of Frankfurtian bullshit (i.e., communications with little to no regard for truth, evidence, or established knowledge). A potential consequence of Frankfurtian bullshit is a deleterious effect on what people perceive to be true. Three experiments test the *insidious bullshit hypotheses* (i.e., bullshitting is evaluated less negatively, but has more negative consequences, than lying) in light of an illusory truth effect (i.e., repeated, false statements are easier to mentally process, and subsequently perceived as more truthful, than new statements). Experiment 1 ($N=204$) analyses revealed that statements from an alleged bullshitter or average person are perceived as equally true, and more true than the very same statements made by a liar. Experiment 2 ($N=293$) and Experiment 3 ($N=318$) revealed bullshitters reliably produce a more extreme illusory truth effect than do objective/honest communicators and liars. Together, these findings suggest that truth-indifferent communication may pose a greater and more insidious threat to public belief than deliberate deception—not because bullshitters intend to mislead, but precisely because they do not.

Keywords Bullshit · Bullshitting · Deception · Lying · Illusory truth · Repetition-induced memory

The global impacts of misinformation and disinformation have continued to attract research attention. The sources of misinformation/disinformation include digital media's engagement-driven newsfeed algorithms, win-at-all-cost tactics in partisan politics, a multitude of significant public crises (Lazer et al., 2018; Lewandowsky et al., 2012, 2017; Vosoughi et al., 2018) and an affinity for *Frankfurtian bullshit* (i.e., communication with little to no regard for truth, established knowledge or genuine evidence; Frankfurt, 1986; Petrocelli, 2018).

Recent research has revealed alarming patterns in how false information spreads in our digital ecosystem. Vosoughi et al. (2018) analyzed over 126,000 news stories shared on Twitter and found that false news reached more people than

true news, with false stories diffusing significantly faster and more broadly than true stories across all categories of information. This spread is facilitated by both technological and human factors. Social media platforms' algorithms often prioritize engaging content over accurate content, while human psychological tendencies—such as confirmation bias and motivated reasoning—make us particularly susceptible to misinformation that aligns with our existing beliefs (Lewandowsky et al., 2017). Among the sources of such misinformation, Frankfurtian bullshit may be uniquely consequential precisely because, unlike deliberate lies, it operates without any regard for truth—making it harder to detect, easier to produce, and, as the present research suggests, more insidious in its effects on what people believe to be true.

Within this broader context of misinformation, bullshitting behavior may be particularly pernicious. Bullshitting behavior may include rhetorical strategies designed to help people appear as if they know what they are talking about in order to impress, persuade, and influence others (Littrell et al. 2021; Petrocelli 2021a), explain things in an area in which their obligations to provide opinions exceeds their

✉ John V. Petrocelli
petrocjv@wfu.edu

¹ Department of Psychology, Wake Forest University, P.O. Box 7778, Winston-Salem, NC 27109, USA

² Stanford University, Palo Alto, CA, USA

actual knowledge in that area (Petrocelli, 2018), or to simply kill awkward silence (Petrocelli 2021b). The contemporary media environment, with its demands for constant content creation and rapid response to events, may create ideal conditions for bullshitting to flourish. Unlike deliberate disinformation, which requires effort to fabricate false claims, bullshit can be produced quickly and effortlessly, making it particularly well-suited to social media's rapid pace and emphasis on continuous engagement.

Although common experience suggests bullshit is ubiquitous, particular social conditions make bullshitting behavior even more likely to occur. For instance, previous empirical research (Petrocelli, 2018; Petrocelli et al., 2020) found that when people feel an obligation to provide an opinion (even if it is an uninformed one), feel knowledgeable on a given topic, or believe that they can “get away” or receive a social pass with their bullshit, bullshitting behavior is more likely to occur.

Given its seemingly ubiquitous status, surprisingly little social science research speaks to the potentially harmful consequences of bullshit on beliefs (Petrocelli, 2018)—and few things are more fundamental to clear reasoning and optimal judgment and decision making than what people believe to be true (Alcock, 2018). Bullshit-based beliefs may vary in their potential harm. For instance, if we believe the claim that putting batteries in the freezer will make them last longer, little harm may come from it. However, if we believe the claim that ingesting disinfectant is an effective way to ward off illness, great harm may ensue.

Central to the potential consequences of bullshit is the *insidious bullshit hypothesis* first proposed by Frankfurt (1986). Frankfurt argued that although common experience suggests bullshitting behavior is evaluated less negatively than lying, bullshit is more insidious in that it leads to more negative consequences to society than lies (see: Kimbrough, 2006; Reisch, 2006). In fact, recent empirical support confirms the first part of Frankfurt's proposal—people do appear to evaluate bullshit less negatively than lying (Petrocelli et al. 2023a, b).

However, the consequences of repeated bullshit exposure are less clear. The current investigation was designed to examine one such consequence, a deleterious effect on what people perceive to be true. Three experiments were conducted to test Frankfurt's (1986) insidious bullshit hypotheses (i.e., bullshitting is evaluated less negatively, but has more negative consequences, than lying) in light of an *illusory truth effect* (i.e., repeated, false statements are easier to mentally process, and subsequently perceived to be more truthful, than new statements; Fazio et al., 2015, 2019; Hasher et al., 1977).

Frankfurtian bullshitting and lying

Because Frankfurtian bullshitting involves communication without regard for truth, it can be difficult to differentiate bullshitting from lying. Importantly, the bullshitter has little to no concern with the truth of what they are communicating. The Frankfurtian bullshitter is often a careless thinker who plays with different ideas in order to advance their own goals, whether that is to sell a car, express a political idea, or replace awkward silence (Littrell et al., 2021; Petrocelli, 2018). Most instances of bullshitting appear to manipulate impressions or opinions (e.g., persuading a person to accept a particular claim). For example, the statement “Our sun is a yellow dwarf star.” can be communicated with little to no concern for the truth, making it bullshit (despite its truth), or can be communicated with great concern for the truth, making it a factual statement (and not bullshit). Therefore, bullshit does not have to necessarily be false—all that is required is that a statement is made with little to no concern for the truth.

With successful lying, an individual must understand the truth, then endorse and work to communicate something other than the truth—the motivation is to get an interlocuter/audience to believe something the liar understands to be false. For it to be a conscious lie, there must be some form of truth, evidence, or any form of established knowledge that an individual knows *before* they spread the lie. The statement “Our sun is a red giant star.” is a lie if it goes against one's knowledge and is made with an intent to deceive. However, it would not be considered a lie, but rather incorrect bullshit if it was communicated with little to no concern for said established knowledge.

Bullshit depends upon the individual's motivation to communicate something without regard for the truth, whereas a lie depends upon the content of the statement and what the communicator believes to be true. However, a commonality of the liar and the bullshitter is that they both present themselves as genuine actors who *appear to be* concerned with the truth (Petrocelli et al. 2023a, b)—but only the liar is actually concerned with truth, so far as to ensure communicating something that is not true (i.e., the lie).

However, it is important to note that lying is not always done for nefarious purposes; otherwise, irony, acting, and sarcasm (“Fantastic, climate change is just what we needed right now.”) and many other forms of entertainment would be considered lying (Mahon, 2008; Rutschmann & Wiegmann, 2017). For this reason, most (but not all) scholars in both psychology and philosophy adhere to the traditional definition of lying that includes the speaker's motivation behind the utterance: A lie is an assertion that the

communicator believes to be false but communicates with the intention to deceive (e.g., Arico & Fallis, 2013; Augustine, 1887; Frankfurt, 1986; Turri & Turri, 2015; Williams, 2002). Thus, similar to bullshit, a lie is not rooted in the truth-value of the assertion, but in the state of mind and intention of the communicator. The liar *intends* to deceive, has regard for the truth, and *believes* the assertion is false, and attempts to *knowingly and intentionally mislead* others (DePaulo et al., 1996; Ekman, 1985; Frankfurt, 1986; Williams, 2002). The bullshitter, by contrast, has no regard for truth or evidence, does not know or care whether what they communicate is true, and need not intend to deceive—yet both the liar and bullshitter present themselves as though they are genuinely concerned with the truth.

Recent empirical work has advanced our understanding of bullshitting behavior beyond theoretical distinctions. Littrell et al. (2021) developed and validated the Bullshitting Frequency Scale, identifying two distinct types of bullshitting: *persuasive bullshitting* intended to impress others, and *evasive bullshitting* aimed at avoiding direct answers or responsibility. This distinction has proven important for understanding how different forms of bullshitting relate to personality traits and cognitive abilities. For instance, Ashton et al. (2023) found that low Honesty-Humility was associated with both types of bullshitting, but showed stronger relationships with persuasive bullshitting.

The cognitive mechanisms underlying bullshit production and detection have also received increased attention. Littrell and Fugelsang (2024) demonstrated that both highly bullshit-receptive and highly bullshit-resistant individuals are poorly calibrated in their ability to detect bullshit, suggesting that traditional explanations focusing solely on analytical thinking may be insufficient. Research has further linked bullshitting to political communication—where it can convey contradictory positions simultaneously (Meibauer, 2021)—and to personality differences in deceptive behavior, with distinct facets of Machiavellianism predicting persuasive versus evasive bullshitting (Blötnér & Bergold, 2023). All of these findings suggest that bullshitting represents a unique form of communication with its own psychological antecedents and mechanisms, distinct from straightforward deception or lying. Critically, this uniqueness has implications not only for how bullshit is produced and detected, but also for how exposure to it may shape beliefs—a question that motivates the current investigation.

Bullshit and perceived truth

How do differences in reactions to bullshitters versus liars alter perceptions about what people believe to be true? We propose that one of the ways in which bullshit may influence

beliefs about what is true, beyond that of lies, is through an illusory truth effect. An illusory truth effect occurs when repeated statements become easier to mentally process and remember. Repeated information tends to be associated with greater processing fluency relative to new statements (Jacoby & Dallas, 1981; Jacoby & Kelley, 1987; Jacoby et al., 1989; Reber & Schwarz, 1999; Unkelbach, 2007), regardless of whether the repeated statements are truthful (Arkes et al., 1989; Bacon, 1979; Begg et al., 1992; Hasher et al., 1977; Johar & Roggeveen, 2007). Initial exposure to information sets a contrast with new information not previously processed, making the previous information relatively easier to recognize (or fluent) than new information, with this increase in processing fluency being confused for truth. Therefore, if an individual reads or hears the statement many times over (e.g., “Our sun is a red giant star.”), they are more likely to believe that statement than a new, yet truthful, statement (e.g., “Our sun is a yellow dwarf star.”). The illusory truth effect is remarkably robust across different types of claims and contexts including trivial statements, consumer products, and political opinions (Arkes et al., 1989; Bacon, 1979; Hawkins & Hoch, 1992; Johar & Roggeveen, 2007). Recent research has demonstrated that repetition increases perceived truth even for highly implausible statements, suggesting the effect is not limited by initial plausibility (Fazio et al., 2019). Even more extreme, the illusory truth effect can emerge even among knowledgeable individuals who should otherwise discern fact from fiction (Fazio et al., 2015). The magnitude of the effect appears to follow a logarithmic pattern, with the largest increase in perceived truth occurring after the first repetition and diminishing returns for subsequent exposures (Hassan & Barber, 2021; see also Lacasagne et al., 2022, for evidence that additional repetitions can extend this effect even to highly implausible content). The timing between exposures also plays a critical role; its magnitude tends to decrease as the delay between repetitions increases (Henderson et al., 2021).

Notably, the effect persists even among knowledgeable individuals (Brashier et al., 2017; Fazio et al., 2019; Wang et al., 2016) and remains robust across differences in cognitive ability, need for cognitive closure, and analytical thinking style (De keersmaecker et al., 2020), suggesting it operates through automatic processing mechanisms largely independent of higher-order cognition. The most comprehensive quantitative synthesis of this literature to date reinforces these conclusions. Ye et al. (2026) conducted a meta-analysis of 182 studies and 366 effect sizes ($N=31,184$ participants), finding a small but robust illusory truth effect after correcting for publication bias. Importantly, their moderator analyses identified the conditions of *initial exposure* as particularly consequential: the presence of explicit veracity cues during first exposure and longer presentation durations

both reduced the magnitude of the effect, suggesting that the circumstances under which people are first exposed to information critically shape how susceptible they later are to repetition-induced truth. These findings provide additional meta-analytic grounding for our prediction that identified bullshit—which characteristically lacks clear veracity signals—should produce especially potent illusory truth effects.

A key question, however, is how Frankfurt's (1986) speaker-level concept—a communicator's indifference to truth—connects to the cognitive mechanisms that produce the illusory truth effect, namely processing fluency and familiarity. We propose that the link is source-mediated: the cognitive impact of repeated exposure depends critically on how recipients evaluate the source of the information. When a communicator is identified as a liar, recipients are motivated to critically scrutinize and reject their claims, a response that disrupts the fluency-to-truth pathway by introducing deliberate, effortful reprocessing. In contrast, because a bullshitter's claims are not categorically false, recipients lack a clear basis for rejection, and repeated exposure proceeds with relatively little critical interference. As a result, the increase in processing fluency that accumulates across repetitions is more likely to be misattributed to truth when claims come from a bullshitter than from a liar. In other words, it is not the bullshitter's indifference to truth per se that drives the effect, but the recipient's corresponding indifference to skepticism—bullshit would bypass the corrective mechanisms that otherwise attenuate the fluency-based illusion of truth. The following five theoretical positions further support this account.

Does the underestimation of negative consequences of bullshit place bullshitters in a more potent position than liars to affect beliefs about what is true? We propose that bullshit can lead to a more potent illusory truth effect relative to that of standard lying and/or honest and trustworthy communications on the basis of five theoretical positions. First, according to Frankfurt's (1986) insidious bullshit hypothesis, bullshit is more insidious than lies because liars are partly concerned with the truth (insofar as telling successful lies) and bullshitters often receive a social pass of acceptance that liars do not. Consistent with general overconfidence and optimism (Fischhoff et al., 1977; Langer & Roth, 1975; Lichtenstein & Fischhoff, 1977, 1980; Koriati et al., 1980), people also tend to think they can effectively detect bullshit—despite the fact that much research suggests they do not (Bond & DePaulo, 2006, 2008; DePaulo et al., 1997; Ekman, 2009; Fisher & Keil, 2016; Kominsky & Keil, 2014; Kruger & Dunning, 1999; Lawson, 2006; Levine & McCornack, 1992; Levine et al., 1999; Littrell & Fugelsang, 2024; Newman et al., 2003; Pennycook et al., 2015; Pennycook & Rand, 2019, 2020; Rozenblit & Keil,

2002; Wilson & Keil, 1998)—and they tend to wrongly assume that bullshit is relatively harmless and has no effects on what they believe to be true.

Second, according to the *dismissal readiness hypothesis* (Petrocelli et al. 2023a, b) people have a tendency to dismiss lies more readily than bullshit because anything categorized as a lie is known to be false, whereas bullshit is not necessarily false. It is with such a difference in the certainty of truth that people should be more likely to reject information from identified liars more readily than bullshitters. When people discover they have been lied to, they know immediately, and with certainty, the information they received is categorically false (to the extent the liar knew the truth to begin with). On the other hand, when people know they have been exposed to bullshit (e.g., it is clear the communicator does not know what they are talking about), people do not know for certain the truth value of the information (Petrocelli et al. 2023a, b). Thus, discovering something is a lie provides greater clarity about the truth than does discovering something is bullshit. Furthermore, feedback regarding one's false beliefs must be immediate and clear in order for corrective efforts to succeed (see: Castelli & Gheiti, 2014; Fazio & Marsh, 2010). The diagnosticity of the truth that comes with discovering lies provides greater certainty of the truth than that provided by ambiguously true/correct or false/incorrect bullshit.

Third, in our investigation we manipulated when people were made aware of the fact that some of the information they were exposed to was false, conscious lies, or bullshit. When given the forewarning they may be exposed to false information, lies or bullshit, people are likely to anticipate the protection of their abilities (though often fallible) to detect it and dismiss it as they “usually” do. However, consistent with compensatory strategies (Martin, 1986; Martin et al., 1990; Wegener & Petty, 1995), when made aware they have unknowingly been exposed to, and potentially contaminated by deceptive information, people should be more motivated to correct for any unwanted, inaccurate information that may have altered their beliefs (Wilson & Brekke, 1994). Because bullshit is not necessarily incorrect, and because people appear to recognize that bullshit, but not the lie, is potentially correct (Petrocelli et al. 2023a, b), perceivers should be less concerned they have been exposed to inaccurate information and therefore be less motivated to correct their perceptions in light of bullshit versus lies.

Fourth, theoretical support for our hypotheses comes from previous research seeking to understand how mental systems come to believe. When people communicate, they operate under the assumption their interlocutors communicate correct information unless given reason to believe otherwise (Grice, 1975). Both the mental systems belief model (Gilbert, 1991; Gilbert et al., 1990) and truth-default theory (Levine, 2014) hold that when individuals initially

comprehend new information, they tend to operate from a “truth-default” mode. Rejection of false information tends to require substantial *type 2 processing* (i.e., deliberate and controlled). However, the relatively few dishonest and untrustworthy behavioral cues necessary to form inductive trait inferences (see: Reeder & Brewer, 1979; Rothbart & Park, 1986) suggest that people are hypersensitive to clear deception signals (e.g., “This guy is a liar.”). Thus, people should be less likely to perceive truth in the claims of those they categorize as liars, who intend to communicate falsehoods, than those they categorize as bullshitters, who may only haphazardly communicate falsehoods.

Finally, several established theories from social and cognitive psychology appear to support why bullshit could be more insidious than lies from the standpoint of an illusory truth effect. While these frameworks were not directly tested in the current investigation, they offer converging theoretical support for why truth-indifferent communication may be especially potent in amplifying the illusory truth effect relative to deliberate lies. For instance, fuzzy-trace theory (Reyna & Brainerd, 2011) suggests that people encode information at two levels: verbatim (exact details) and gist (general meaning). Bullshit’s strategic ambiguity may exploit this dual-processing system by making it difficult to encode precise verbatim traces while simultaneously creating vague but persuasive gist representations. In contrast, lies typically provide clear verbatim information that can be more readily evaluated and potentially rejected. The continued influence effect (CIE; Lewandowsky et al., 2012) demonstrates that misinformation can persist in memory even after explicit corrections. The CIE may be particularly pronounced with bullshit because its inherent ambiguity makes it more resistant to direct refutation than lies. While lies can be definitively proven false, bullshit’s indifference to truth creates what Pennycook and Rand (2019) term a ‘meta-confusion’ about the very nature of truth and evidence. Finally, research on epistemic vigilance (Sperber et al., 2010) suggests that humans have evolved mechanisms to evaluate both the content of communications and the trustworthiness of sources. Bullshit may exploit a gap in these defense mechanisms: while we are well-adapted to detect deliberate deception (lies), we may be less equipped to handle communication that is indifferent to truth altogether. This aligns with recent work showing that people are generally better at detecting lies than identifying bullshit (Littrell & Fugelsang, 2024), perhaps because our epistemic vigilance systems evolved primarily to combat intentional deception rather than truth-indifferent communication.

Overview of experiments

Three experiments were designed to test the insidious bullshit hypothesis through a traditional illusory truth effect procedural paradigm. The purpose of Experiment 1 was to directly test the possibility that the perceived truth values of both true and false claims made by bullshitters are greater than those of the very same claims made by liars. Therefore, in Experiment 1, we manipulated the source of true and false claims (i.e., average person, bullshitter, liar) and asked participants to report how true they perceived the claims to be. We hypothesized people find bullshitter claims to be more true (less false) than liar claims. If people do perceive a bullshitter’s claims to be more true (or less false) than that of a liar’s, the consequences to beliefs following repeated exposure to a bullshitter’s claims may be more consequential than repeated exposure to a liar’s claims. Experiment 2 and Experiment 3 tested this possibility using a traditional illusory truth effect procedural paradigm. Therefore, in Experiment 2 and Experiment 3, we manipulated the relative exposure frequency of true and false claims and again manipulated the source. Consistent with several prior demonstrations of the illusory truth effect (Fazio et al., 2015, 2019), we hypothesized that more pronounced illusory truth effects would be found with increasing exposure to the claims and especially when the claims were allegedly made by a source instructed to bullshit relative to a source instructed to tell the truth or lie.

In Experiment 2 and Experiment 3, we also manipulated *when* participants became privy to the fact they would be, or were, provided with some information that was, or would be, lies or bullshit. Consistent with prior demonstrations suggesting that discounting cues (e.g., the source is a liar or bullshitter) affect judgments most profoundly when the discounting cues come after initial exposure to false information (Foos et al. 2016; Petrocelli et al. 2023a, b), we hypothesized the classic illusory truth effect would emerge when the discounting cue (i.e., the source was instructed to provide objectively false information, lie or bullshit) preceded initial exposure to the claims. However, because bullshit as a discounting cue carries less weight with regard to falsity than does objectively false information or lies, we expected the illusory truth effect to be sustained in the bullshit condition, but attenuated in the objectively false and lie conditions, when the discounting cues followed initial exposure to the claims. Such a pattern of findings would serve as the very first demonstration of support for the insidious bullshit hypothesis through a pronounced illusory truth effect for bullshit relative to lies.

Experiment 1

In Experiment 1, we asked participants to rate the truthfulness of a series of statements (half true and half false) allegedly made by one of three individuals. Each of the individuals were allegedly described by his friends, family, and colleagues as either an average person, a liar or a bullshitter. Because people tend to be more critical of liars than bullshitters (Petrocelli et al. 2023a, b), we hypothesized that our participants would rate the true value of bullshitters to be greater than that of liars. Experiment 1 permitted the test of the hypothesis that people perceive significantly less truth from the claims of liars than they do from the claims of bullshitters. In summary, the primary hypothesis for Experiment 1 was: statements attributed to a bullshitter would be rated as truer than the identical statements attributed to a liar, regardless of whether those statements were actually true or false.

Method

Participants and design

A sample of 204 university undergraduates (65.70% female, $M_{\text{age}} = 18.71$, $SD_{\text{age}} = 0.89$), enrolled in an introductory psychology course, participated in exchange for partial course credit.

The experiment employed a 3 (Author Description: liar, bullshitter, control) $\times$ 2 (Statement: true vs. false) mixed factorial design with the latter variable being the within-subjects variable and Truth Rating as the dependent variable. Critically, the statements themselves were held constant across all three Author Description conditions; only the description of the source varied. This was a deliberate design choice—by keeping statement content identical, any differences in Truth Ratings across conditions can be attributed specifically to participants' perceptions of the source rather than to differences in the content or plausibility of the statements themselves, thereby isolating source attribution as the active variable.

A sensitivity power analysis using G*Power 3.1.9.6 (Faul et al., 2007) was conducted for the main effect between the experimental manipulation of Author Description, in a mixed analysis of variance (ANOVA) test with three between-subjects conditions (representing the experimental conditions, the within-subjects factor, and two-way interaction term). The analysis indicated that with a significance level of $p = .05$ (two-tailed), an expected correlation among the within-subjects measurements of $r = .50$, and a power of $\beta = 0.80$, the final sample size provided sufficient power to detect an effect with $f = 0.11$, corresponding to an $\eta^2_p = 0.01$

(small). We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data and procedures are available online (<https://osf.io/3syxn/overview>) and no data were collected after data analysis.

Procedure

All experimental materials were presented through a self-administered computer questionnaire using MediaLab v2016 Research Software (Jarvis, 2016); participants advanced by clicking appropriate response keys. Participants were informed they would be presented with various statements and that we were interested in surveying social perceptions of the truth of these statements.

Author description manipulation Before participants were presented with the list of statements, it was explained that the authors of the statements had been instructed to compile a list of several claims on the basis of their own knowledge. Participants were then randomly assigned to receive one of three sets of information about the author of the statements they were asked to evaluate for truth. All participants read: "Please understand that the entire list of statements we will present to you are from an author named Jim Smith. Jim Smith is described by his friends, family, and colleagues as a ()." Participants assigned to the *control condition* were led to believe the author of the statements is described by his friends, family, and colleagues as an "average person", whereas participants assigned to the *liar condition* were led to believe the author of the statements is described by his friends, family, and colleagues as a "liar". Participants assigned to the *bullshitter condition* were led to believe the author is described by his friends, family, and colleagues as a "bullshitter". Because bullshitting is commonly confused with lying, in the case of the bullshitter condition it was further explained that a bullshitter says things without regard for truth, genuine evidence, or established knowledge, nor do they care to know if what they say is true (Frankfurt, 1986; Petrocelli, 2018).

Attention check Just before being presented with the list of statements, participants were queried with an attention check including the following: "The author of the list of statements we will present to you are from an author named Jim Smith. Jim Smith is described by his friends, family, and colleagues as a(n) ____." Participants selected from one of nine options that included "Liar", "BSer" and "Average Person" among six additional, filler options (e.g., "Great Doctor"), and permitted to click a Go Back button (which took them back to the previous author description screen frame) if they were uncertain. The total correct response

rate was 97.05% and did not depend on Author Description condition.

Truth ratings task Participants rated their perceived truth of 42 statements randomly sourced from Tauber et al. (2013), half of which were true (e.g., “Helsinki is the capitol of Finland.”) and half of which were false (e.g., “Bell is the inventor of the wireless radio.”). Truth Ratings in response to “How much do YOU believe the following statement is likely to be true?” were made using an 11-point response scale with *not at all likely to be true* (0) and *extremely likely to be true* (10) as the anchor labels. Following the truth ratings task, participants were thanked for their participation and dismissed.

Results and discussion

Truth Rating data were subjected to a 3 (Author Description: liar, bullshitter, control) $\times$ 2 (Statement: true vs. false) repeated measures ANOVA. As would be expected from prior research (see: Fazio et al., 2015), participants successfully rated true statements as significantly more truthful ($M=6.36$, $SD=1.39$) than they did false statements ($M=6.00$, $SD=1.34$), $F(1, 201)=28.81$, $p < .001$, $\eta^2_p = 0.13$. Most importantly, the Author Description main effect emerged as statistically significant, $F(2, 201)=11.34$, $p < .001$, $\eta^2_p = 0.10$. As expected, follow-up contrasts revealed the overall average Truth Ratings for statements derived from the control condition source ($M=6.65$, $SD=1.02$) were significantly greater than both those allegedly derived from a bullshitter ($M=6.23$, $SD=1.00$), $t(201)=3.68$, $p < .001$, $d = 0.42$, and allegedly derived from a liar ($M=5.65$, $SD=1.57$) did differ significantly from those of the control condition, $t(201)=8.64$, $p < .001$, $d = 0.99$. Consistent with our hypotheses, the mean Truth Rating for statements allegedly derived from a bullshitter were significantly greater than those allegedly derived from a liar, $t(201)=5.10$, $p < .001$, $d = 0.58$. These results were not further qualified by the Author Description $\times$ Statement interaction, $F(1, 201) = 0.24$, $p = .787$.

These results support the conclusion that people find the truth of what bullshitters communicate to be truer than when liars make the very same claims. In each context, whether the statements are actually true or false, our participants rated the truth of statements allegedly made by bullshitters to be more true than the very same statements allegedly made by liars, though less true than they perceived the very same statements allegedly made by “average” people. These results satisfy a necessary condition for the validity of our theoretical position that a potential illusory truth effect through bullshitting may be more potent than that which can

emerge through lying, if in fact people are to treat the truth likelihoods of claims made by bullshitters to be significantly greater than those of the same claims made by liars. These results not only support Frankfurt’s distinction between bullshit and lies but also establish a crucial precondition for testing his insidiousness hypothesis—if bullshit statements are perceived as more credible than lies from the outset, this opens the possibility for bullshit to have more pernicious effects through repeated exposure. Thus, the stage is set for testing whether this initial credibility advantage translates into stronger illusory truth effects in subsequent experiments. Experiment 2 and Experiment 3 were designed to test this possibility.

Experiment 2

Having established through Experiment 1 that when people understand what Frankfurtian bullshit (i.e., does not regard truth and could sometimes be true) is and how it differs from lying, they expect/perceive bullshit to be more true (or less false) than lies. An essential difference between bullshitting and lying as it pertains to beliefs in the truth of a claim is that when these behaviors occur equally frequently social observers can expect the base rate accuracy of bullshitters to be greater than that of liars. We are interested in how these base rates of expected accuracy impact what people believe to be true in the context of repeated exposure.

In Experiment 2, we again asked participants to rate the truthfulness of a series of statements (half true and half false) allegedly made by one of three individuals allegedly serving as authors. Participants were informed that some of the statements were true, but that each of the authors were also provided with instructions by the experimenters to generate either objectively false information, lies, or bullshit for some of their statements. Consistent with a standard illusory truth procedural paradigm, we also varied the frequency of exposure to some of the claims. Given the base rate accuracy of bullshit should be greater than that of direct intentions to deliver false information, we expected the bullshitter condition to produce a more pronounced illusory truth effect than the object/honest and liar conditions. In Experiment 2, we also manipulated when participants became privy to the fact they would be, or were, provided with some information that was, or would be, objectively false, lies or bullshit. Given discounting cues (e.g., the source is a liar or bullshitter) tend to affect judgments profoundly only when the discounting cues come after initial exposure to false information (Foos et al. 2016; Petrocelli et al. 2023a, b), we hypothesized the classic illusory truth effect would emerge when the discounting cue (i.e., the source was instructed to provide objectively false information, lie or bullshit) preceded initial exposure

to the claims. However, because bullshit as a discounting cue carries less weight with regard to falsity than does objectively false information or lies, we expected the illusory truth effect to be sustained in the bullshit condition, but attenuated in the objectively false and lie conditions, when the discounting cues followed initial exposure to the claims. Importantly, in Experiment 2, we operationalized Frankfurt's (1986) concept of insidiousness through the magnitude of the illusory truth effect. Specifically, if bullshit is more insidious than lying, it should produce a stronger illusory truth effect—that is, repeated bullshit should increase perceived truth more than repeated lies. This operationalization allows us to directly test Frankfurt's insidious bullshit hypothesis by measuring whether bullshit's influence on truth judgments grows more substantially with repetition compared to lies or objectively false statements. It should be noted that this operationalization captures one specific cognitive manifestation of insidiousness—its effect on repetition-based truth perception—rather than the full scope of Frankfurt's broader philosophical argument about the corrosive role of bullshit in public discourse. The present experiments are therefore best understood as an empirical test of one measurable consequence of truth-indifferent communication, not a comprehensive evaluation of Frankfurt's thesis. In summary, the primary hypotheses for Experiment 2 were: (1) repeated exposure to bullshitter-sourced claims would produce a more pronounced illusory truth effect than repeated exposure to liar- or control-sourced claims; and (2) this effect would be strongest—and most resistant to correction—when participants were informed about the nature of the source only after initial exposure to the statements.

Method

Participants and design

A sample of 293 university undergraduates (63.50% female, $M_{\text{age}} = 18.43$, $SD_{\text{age}} = 0.74$), enrolled in an introductory psychology course, participated in exchange for partial course credit.

The experiment employed a 3 (Author: objective/honest (control), liar, bullshitter) $\times$ 2 (Informed About Falsity: before vs. after initial exposure) $\times$ 2 (Statement: true vs. false) $\times$ 3 (Exposure: 1, 2, 3) mixed factorial design with the latter two variables being within-subjects variables and Truth Rating as the dependent variable.

A sensitivity power analysis using G*Power 3.1.9.6 (Faul et al., 2007) was conducted for the interaction effect between the experimental manipulations of Author Description and Exposure in a mixed ANOVA. The analysis indicated that with a significance level of $p = .05$ (two-tailed), an expected correlation among the within-subjects measurements of r

$= .50$, and a power of $\beta = 0.80$, the final sample size provided sufficient power to detect an effect with $f = 0.08$, corresponding to an $\eta^2_p = 0.02$. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study. All data and procedures are available online (<https://osf.io/3syxn/overview>) and no data were collected after data analysis.

Procedure

All experimental materials were presented through a self-administered computer questionnaire using MediaLab v2016 Research Software (Jarvis, 2016); participants advanced by clicking appropriate response keys. The procedures followed the basic guidelines of a traditional illusory truth procedural paradigm employed by Fazio et al. (2015) with two additional manipulations (i.e., author manipulation and when informed about falsity).

Initial exposure and interest rating task First, participants were presented with 66 trials of statements in which they were only asked to rate their perceptions of how interesting each statement was on a 6-point response scale with *very interesting* (1) to *very uninteresting* (6) as the anchor labels. Across the trials, participants were presented with 44 unique statements whereby 22 were displayed once and 22 were displayed twice. Half of the statements were true and half were false. The primary purpose of this task was to set an initial exposure of content that would set a contrast with newly presented information, not previously processed, such that the previously presented (repeated) information was relatively easier to recognize (or fluent) than the new information.

Author manipulation Participants were randomly assigned to one of three conditions pertaining to the alleged instructions given to the alleged authors of the statements. Those assigned to the *objective/honest (control) condition* were informed the author of the statements was instructed to compile a list of statements and to "...please know that some statements are in fact true and others are in fact false." Participants assigned to the *liar condition* read the author of the statements was "...instructed to include some statements that *the author knows* to be clearly true and some statements that *the author knows* to be clearly false." Participants assigned to the *bullshitter condition* read the author of the statements was "...instructed to include some statements that *the author knows* to be clearly true and some statements that *the author does not know (nor care)* to be true or false."

Informed about falsity Participants were randomly assigned to one of two conditions pertaining to learning about the falsity of some information within the Initial

Exposure and Interest Rating Task. Half of the participants were informed that some of the information they *were about to be asked about* was false *before beginning* the Initial Exposure and Interest Rating Task, whereas the other half were informed that some of the information they *had already been asked about* was false *after completing* the Initial Exposure and Interest Rating Task.

Truth rating task Finally, participants were presented with 88 trials of statements in which they were asked to rate how true they believed each statement to be on a 6-point response scale with *definitely false* (1) to *definitely true* (6) as the anchor labels. Across the trials, participants were presented with 44 unique statements (not presented during the Initial Exposure and Interest Rating Task), and 44 repeated statements, which had been presented during the Initial Exposure and Interest Rating Task (22 displayed once prior, 22 displayed twice prior). In each of the unique and repeated statement sets, half of the statements were true and half were false (Tauber et al., 2013). Consistent with prior demonstrations of the illusory truth effect (e.g., Fazio et al., 2015), the more difficult items identified by Tauber et al. (2013) were selected to help ensure the emergence of the effect; the full set of 88 items hold a recall rate < 13.0% and a mean recall rate of 5.0%. Following the truth ratings task, participants were thanked for their participation and dismissed.

Results and discussion

Truth Rating data were subjected to a 3 (Author: objective/honest (control), liar, bullshitter) $\times$ 2 (Informed About Falsity: before vs. after initial exposure) $\times$ 2 (Statement: true vs. false) $\times$ 3 (Exposure: 1, 2, 3) repeated measures ANOVA. Consistent with expectations, a statistically significant main effect of Author emerged, $F(2, 287) = 15.18, p < .001, \eta^2_p = 0.10$. As expected, participants in the bullshitter condition reported significantly greater truth ($M = 4.14, SD = 0.35$) than their liar condition counterparts ($M = 3.92, SD = 0.34$), $t(287) = 2.95, p = .003, d = 0.63$, and their control condition counterparts ($M = 3.88, SD = 0.41$), $t(287) = 3.56, p < .001, d = 0.68$; liar and control condition participants did not differ significantly in their overall average Truth Rating, $t(287) = 0.60, p = .548$. The failure to find a difference in overall Truth Rating between the control and liar conditions (like that reported in Experiment 1) can only be explained by the inclusion of repeated exposure and the fact that half of the participants were informed about falsity after initial exposure (not included in Experiment 1). Repeated exposure tended to increase Truth Rating for all conditions (as would be expected of the illusory truth effect), and being informed about falsity after initial exposure tended to reduce Truth

Rating only for the control and liar conditions. Furthermore, Truth Rating was significantly greater when participants learned about falsity within the statements before initial exposure during the interest rating task ($M = 4.08, SD = 0.40$) than when they learned about it after initial exposure ($M = 3.89, SD = 0.35$), $F(1, 287) = 21.23, p < .001, \eta^2_p = 0.07$. Consistent with Experiment 1, and earlier findings (Fazio et al., 2015), a main effect of Statement emerged, such that true statements were rated as significantly more true ($M = 4.06, SD = 0.39$) than were false statements ($M = 3.91, SD = 0.44$), $F(1, 574) = 66.14, p < .001, \eta^2_p = 0.19$. Also as would be expected given several previous illusory truth reports, a main effect of Exposure emerged, $F(2, 574) = 208.65, p < .001, \eta^2_p = 0.42$; Truth Ratings increased significantly when the statement was repeated once ($M = 3.98, SD = 0.56$) relative to newly presented statements ($M = 3.75, SD = 0.36$), $t(574) = 8.92, p < .001$, as well as when the statement was repeated twice ($M = 4.22, SD = 0.60$), relative to when it was repeated once, $t(574) = 9.32, p < .001, d = 0.41$.

These results were qualified by a significant four-way Author $\times$ Informed About Falsity $\times$ Statement $\times$ Exposure interaction, $F(4, 574) = 2.86, p = .023, \eta^2_p = 0.02$. Because the true statement data produced a similar, but less pronounced, pattern found for the false statement data, we simplified our analyses by focusing only on the false statement data along the significant three-way Author $\times$ Informed About Falsity $\times$ Exposure interaction, $F(4, 574) = 3.80, p = .005, \eta^2_p = 0.03$ (see Fig. 1). The key pattern driving this interaction can be summarized as follows: when participants were informed about falsity before initial exposure, Truth Ratings for false statements declined with repeated exposure in the control condition but showed little change in the liar and bullshitter conditions, suggesting that forewarning attenuated the illusory truth effect across all source conditions. In contrast, when participants were informed about falsity only after initial exposure, Truth Ratings for false statements increased substantially with repeated exposure in the bullshitter condition but not in the control or liar conditions—indicating that only bullshitter-sourced claims continued to gain perceived truth after post-hoc disclosure of potential falsity. Pairwise contrast analyses showed when the fact that some of the statements were false was revealed *before* initial exposure, the classic illusory truth effect was observed in the control condition, $t(574) = -5.06, p < .001, d = -0.54$ but neither the liar, $t(574) = -1.88, p = .059$, nor the bullshitter conditions, $t(574) = -1.32, p = .185$. However, when the fact that some of the statements were false was revealed *after* initial exposure, the classic illusory truth effect was observed only in the bullshitter condition, $t(574) = -5.42, p < .001, d = -0.71$, but neither the control, $t(574) = 0.81, p = .414$, nor the liar conditions, $t(574) = 0.31, p = .753$. From another important angle among the informed

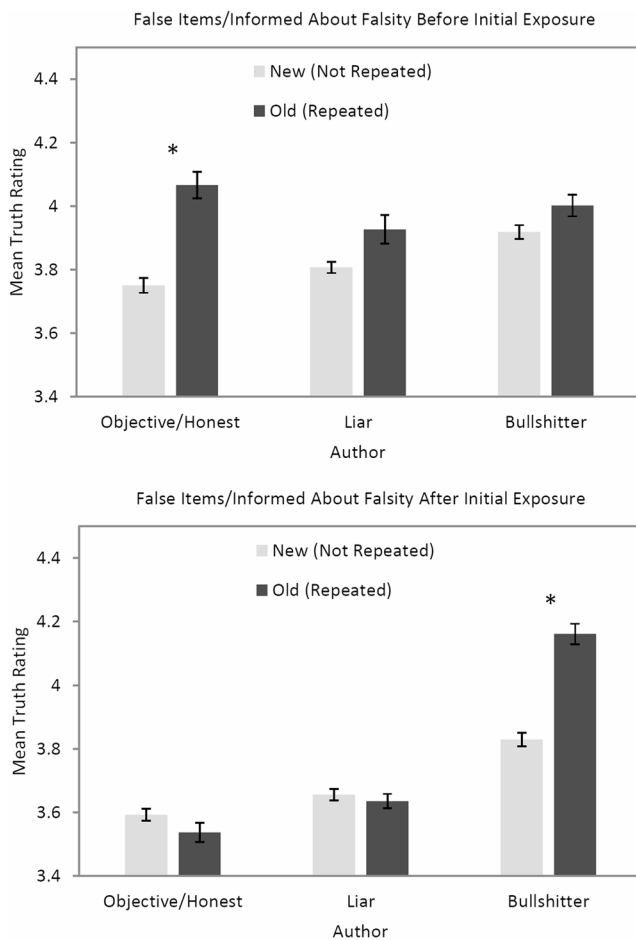


Fig. 1 Mean truth rating for false items by author, informed about falsity, and exposure conditions (Experiment 2). asterisks indicate statistically significant differences in mean truth ratings between adjacent exposure levels within each author condition

about falsity after initial exposure data, repeated exposure resulted in pronounced truth ratings among the bullshitter condition relative to both the control, $t(574)=7.96, p < .001, d=1.16$, and liar conditions, $t(574)=6.78, p < .001, d=1.09$; whereas no difference was observed between the control and liar conditions, $t(574)=1.23, p = .219$.

Our Experiment 2 results further support the insidious bullshit hypothesis. A more potent illusory truth effect appeared for bullshit more than it did for the same content pitched as potential lies and generally honest/truthful information when participants were made aware of the nature of some statements (i.e., bullshit or falsity) after initial exposure. Whereas participants appeared to “correct” their judgments of truth when they were unknowingly exposed to false information (i.e., through the control and liar conditions), such a correction was not observed when they were unknowingly exposed to potentially false information of which even the alleged author was not fully aware of, nor

concerned with, the truthiness of the claims (i.e., bullshitter condition).

The emergence of a more pronounced illusory truth effect for bullshit compared to lies directly supports Frankfurt’s (1986) insidiousness hypothesis. By operationalizing insidiousness through the magnitude of the illusory truth effect, we can quantifiably demonstrate that bullshit is indeed more harmful than lying—it produces stronger truth effects through repetition. This is particularly evident in the condition where participants learned about the nature of the statements after initial exposure, suggesting that bullshit’s insidiousness may be most powerful when people are not initially primed to be skeptical.

One of the disadvantages of Experiment 2 is that participants may not have clearly understood that regardless of the alleged author instructions, *exactly half* the statements they evaluated were true and half were false. Such ambiguity may have unduly affected Truth Ratings given that liars may not be expected to lie as much as bullshitters bullshit because the social consequences of lying are often greater than that of bullshitting (Petrocelli et al. 2023a, b). Therefore, Experiment 3 was designed to test if our Experiment 2 findings could be replicated in the context of equally frequent provision of objectively false information, lying and bullshitting.

Experiment 3

Experiment 3 followed the very same procedures as those employed in Experiment 2 with one exception to the description of the instructions provided to the alleged authors of the statements. Participants were informed that exactly half of the statements they evaluated were objectively false, lies or bullshit. Doing so permits attributing any differences in the illusory truth effect to the base rate accuracy associated with objectively false information, lies and bullshit, and not to any perceived differences in the rate of the behaviors. We again expected bullshit to result in a more pronounced illusory truth effect especially when the discounting cue was delivered only after initial exposure to the repeated claims. Following the same operationalization of insidiousness through the magnitude of the illusory truth effect used in Experiment 2, we again tested whether bullshit produces stronger truth effects than lies, this time controlling for the frequency of false claims. In summary, the primary hypothesis for Experiment 3 mirrored that of Experiment 2: bullshitter-sourced claims would produce a stronger illusory truth effect than liar- or control-sourced claims, even when the frequency of false statements was held equal across conditions.

Method

Participants and design

A sample of 318 university undergraduates (59.40% female, $M_{\text{age}} = 18.40$, $SD_{\text{age}} = 1.87$), enrolled in an introductory psychology course, participated in exchange for partial course credit.

The experiment employed the very same structural design as that employed in Experiment 2. Again, a sensitivity power analysis indicated that with a significance level of $p = .05$ (two-tailed), an expected correlation among the within-subjects measurements of $r = .50$, and a power of $\beta = 0.80$, the final sample size provided sufficient power to detect an effect with $f = 0.07$, corresponding to an $\eta^2_p = 0.02$. All open science reporting practices described in Experiments 1 and 2 were followed identically, with data and procedures available at the same OSF link provided above.

Procedure

The materials and procedures were identical to those employed in Experiment 2 with one exception. Once again, participants were randomly assigned to one of three conditions pertaining to the alleged instructions given to the alleged authors of the statements. However, in Experiment 3, participants assigned to the *objective/honest (control) condition* were informed the author of the statements was instructed to compile a list of statements. It was then explained that we determined some of the statements to be true/correct and some statements to be false/incorrect and sorted them into two separate categories. Finally, it was explained that the list of statements included both true and false claims, and exactly half of the statements were true and half of the statements were false.

Participants assigned to the *liar condition* were informed that the author was instructed to write some statements they knew to be true/correct as well as statements they *knew to be false/incorrect*. In fact, for these latter statements, the authors were instructed to write statements *they knew to be lies*. For each statement authored, the author also indicated whether it was a statement they knew to be true or a statement they *knew to be false/lies*. Participants were then led to believe we sorted the statements into two separate categories, and that the list of statements presented to them included both true and *false/lie statements*. It was explained that exactly half of the statements were true and half of the statements were *false/lie statements*.

Participants assigned to the *bullshitter condition* were informed that the author was instructed to write some statements they knew to be true/correct as well as statements they *did not know (nor care) to be true/correct or false/*

incorrect. In fact, for these latter statements, the authors were instructed to write statements *without any real concern for facts, truth, genuine evidence, or existing knowledge*. For each statement authored, the author also indicated whether it was a statement they knew to be true or a statement they *did not know, nor care, if it was true (i.e., disregard-for-truth statements)*. Participants were then led to believe we sorted the statements into two separate categories, and that the list of statements presented to them included both true and *disregard-for-truth statements*. It was explained that exactly half of the statements were true and half of the statements were *disregard-for-truth statements*.

All other procedures were identical to Experiment 2. Following the truth ratings task, participants were thanked for their participation and dismissed.

Results and discussion

Truth Rating data were subjected to a 3 (Author: objective/honest (control), liar, bullshitter) $\times$ 2 (Informed About Falsity: before vs. after initial exposure) $\times$ 2 (Statement: true vs. false) $\times$ 3 (Exposure: 1, 2, 3) repeated measures ANOVA. Consistent with expectations, a statistically significant main effect of Author emerged, $F(2, 312) = 13.97$, $p < .001$, $\eta^2_p = 0.08$. As expected, participants in the bullshitter condition reported significantly greater truth ($M = 4.10$, $SD = 0.36$) than their liar condition counterparts ($M = 3.93$, $SD = 0.35$), $t(312) = 2.25$, $p = .024$, $d = 0.47$, and their control condition counterparts ($M = 3.89$, $SD = 0.38$), $t(312) = 2.85$, $p < .004$, $d = 0.56$; liar and control condition participants did not differ significantly in their overall average Truth Rating, $t(312) = 0.61$, $p = .545$. Furthermore, Truth Rating was significantly greater when participants learned about falsity within the statements before initial exposure during the interest rating task ($M = 4.03$, $SD = 0.34$) than when they learned about it only after initial exposure ($M = 3.91$, $SD = 0.39$), $F(1, 312) = 13.98$, $p < .001$, $\eta^2_p = 0.04$. Consistent with earlier our Experiment 1 and Experiment 2 findings, a main effect of Statement emerged, such that true statements were rated as significantly more true ($M = 4.03$, $SD = 0.33$) than were false statements ($M = 3.91$, $SD = 0.39$), $F(1, 624) = 49.72$, $p < .001$, $\eta^2_p = 0.14$. A main effect of Exposure emerged, $F(2, 624) = 224.14$, $p < .001$, $\eta^2_p = 0.41$; Truth Ratings increased significantly when the statement was repeated once ($M = 3.99$, $SD = 0.51$) relative to newly presented statements ($M = 3.73$, $SD = 0.35$), $t(624) = 9.68$, $p < .001$, $d = 0.59$, as well as when the statement was repeated twice ($M = 4.20$, $SD = 0.55$), relative to when it was repeated once, $t(624) = 8.52$, $p < .001$, $d = 0.39$.

The results beyond these main effects were similar to those observed in Experiment 2. Once again, of key interest

was the false statement data along the three-way Author $\times$ Informed About Falsity $\times$ Exposure interaction—however, this interactive effect failed to reach statistical significance, $F(2, 312) = 0.71, p = .493$. Most critical to our analysis, a statistically significant Author $\times$ Exposure interaction emerged, $F(2, 312) = 5.08, p = .007, \eta^2_p = 0.03$ (see Fig. 2). The pattern driving this interaction is straightforward: Truth Ratings for false statements increased as a function of repeated exposure in both the control and bullshitter conditions, but the increase was substantially larger in the bullshitter condition than in either the control or liar conditions. The liar condition showed little increase in perceived truth with repetition, consistent with the view that identifying a source as a liar attenuates the fluency-to-truth pathway. Pairwise contrast analyses showed the classic illusory truth effect was observed in the control condition, $t(312) = -2.97, p = .003, d = -0.33$, and pronounced even more so in the bullshitter condition, $t(312) = -6.14, p < .001, d = -0.68$, but not observed in the liar condition, $t(312) = -1.78, p = .074$. From another important angle, repeated exposure resulted in pronounced truth ratings among the bullshitter condition relative to both the control, $t(312) = 6.86, p < .001, d = 0.64$, and liar conditions, $t(312) = 6.22, p < .001, d = 0.59$; whereas no difference was observed between the control and liar conditions, $t(312) = 0.65, p = .510$.

Thus, similar to Experiment 2, our Experiment 3 results further support the insidious bullshit hypothesis. A more potent illusory truth effect appeared for bullshit more than it did for the same content pitched as lies and generally honest/truthful information. Even when controlling for the frequency of false claims, bullshit continued to produce stronger illusory truth effects than lies. These results provide further evidence that the insidiousness of bullshit, as measured through its ability to generate stronger truth

effects with repetition, is not simply a function of exposure frequency but rather reflects something more fundamental about how bullshit influences belief formation. The fact that this pattern held even with equal rates of false information suggests that bullshit's greater capacity to distort perceived truth is inherent to its nature as truth-indifferent communication rather than being an artifact of how frequently it occurs.

General discussion

Following from Experiment 1, while individuals in their day-to-day experience may not differentiate between bullshit and lies in a rigorous manner, the bullshit-lie distinction is psychologically critical. If people process statements made by bullshitters as more subjectively true relative to the very same statements made by liars, then it is reasonable to conclude that bullshit can be more insidious than lying. Results of Experiment 2 and Experiment 3 also support the insidious bullshit and dismissal hypotheses suggesting that bullshit can produce a stronger illusory truth effect, relative to lies—primarily once people become privy to the fact their initial exposure to semantic content was in fact bullshit. These results suggest that bullshit, more so than lying, can have a deleterious effect on what people perceive to be true—and what people believe to be true is fundamental to clear reasoning and optimal judgment and decision making (Alcock, 2018). Given the effect size demonstrated in these studies, if bullshit can be a dramatically more harmful form of communication compared to lying, then we must strive to understand its impact on optimal judgment and decision making.

The clarifications we included with the instructions allegedly given to our alleged bullshitters in Experiment 2 and Experiment 3 suggest the labels “bullshitter” and “liar” are less important than the critical difference between bullshit and lies as they pertain to truth value—that being the base rate accuracy of truth for bullshit is typically greater than that of lies. We suspect, had we simply used the labels “bullshitter” and “liar”, we would not have obtained the differences in the illusory truth we observed given that these labels are commonly used to mean the same thing (Petrocelli 2021b).

Interestingly, the fact that the bullshit-induced illusory truth effect appears to be significantly more pronounced when people are provided discounting cues *after* initial exposure to statements that may be false (Experiment 2) is consistent with the same before/after pattern of findings found for *sleeping effect* (a persuasive influence that increases, rather than decays, over time; Albarracín et al., 2017; Cook et al., 1979; Hovland & Weiss, 1951; Pratkanis et al., 1988; Priester et al., 1999). Sleeping effects reliably

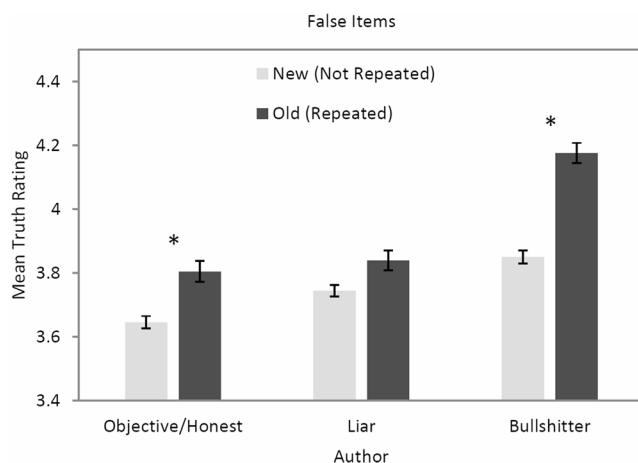


Fig. 2 Mean truth rating for false items by author and exposure conditions (Experiment 3). Asterisks indicate statistically significant differences in mean truth ratings between adjacent exposure levels within each author condition

emerge only when the discounting cue (i.e., notification that some of the information presented earlier may not be entirely correct) comes after (but not before) the initial attitudinal relevant information (Foos et al. 2016; Petrocelli et al. 2023a, b). Typically, a communication has greater impact when attitudes are measured closer, versus later, from the time of the communication. The sleeper effect shows just the opposite pattern, demonstrating higher levels of persuasion when measured following temporal delays. Similar to the current investigation, the Petrocelli et al. (2023a, b) data revealed that downward adjustments to attitudes observed when the discounting cue came *before* exposure to the advertisement claims was not nearly as pronounced as the downward adjustments to attitudes observed when the discounting cue came *after* exposure to the claims. In other words, the current investigation adds to the growing number of demonstrations suggesting that immediate corrective efforts in the face of disinformation are insufficient when disinformation is *forewarned*, relative to when people are made *aware of disinformation exposure only after they have been exposed*. Both illusory truth and sleeper effect procedural paradigms further suggest that if such disinformation comes in the form of bullshit, corrections to beliefs and attitudes are generally insufficient relative to those same claims made in the form of lies.

It is also useful to consider the insidious bullshit hypothesis in light of social judgement theory (Sherif & Hovland, 1961; Sherif & Sherif, 1967; Sherif et al., 1965). Consistent with the dismissal readiness hypothesis, lies are categorically false and thus more readily trigger a receiver's latitude of rejection, whereas bullshit—being ambiguously true or false—may pass through without triggering the same critical response. As a result, repeated bullshit may accumulate within one's latitude of noncommitment, gradually shifting perceived truth in ways that outright lies do not.

Our findings also have important implications for addressing deceptive communication in digital media and public discourse. The discovery that bullshit produces stronger illusory truth effects than lies suggests that fact-checking efforts may need different approaches for these distinct forms of deception. While lies can be directly fact-checked against evidence, bullshit's indifference to truth may require developing new intervention strategies. This concern is underscored by Lee and Bissell's (2024) finding that repeatedly presenting false claims within corrective fact-checking messages can paradoxically increase their perceived credibility—an effect that was strongest among those initially skeptical of the misinformation, suggesting that bullshit's repetition-driven influence may be especially potent for those who would otherwise be inclined to reject it. The practical implication is sobering: standard debunking strategies, designed with deliberate lies in mind, may

be insufficient—and even counterproductive—when the target is truth-indifferent communication. For instance, social media platforms might need different detection algorithms and warning labels for content that displays characteristics of bullshit versus outright falsehoods. Similarly, media literacy programs might benefit from teaching people to distinguish between these different forms of deceptive communication, as the skills needed to detect and resist bullshit appear to differ from those needed to identify lies. Our experimental results also suggest that public policy efforts to combat misinformation should consider whether existing regulatory frameworks, typically designed to address deliberate falsehoods, are adequate for dealing with the more insidious effects of bullshit. Future research might productively examine how institutional policies and educational interventions could be tailored to address the unique challenges posed by truth-indifferent communication.

Our participants were directly informed that the alleged authors of the statements either communicated bullshit or lies; this was necessary to directly test the influence of bullshit and lies on perceived truth. Of course, in practice, people are rarely directly informed they have been exposed to such deception, making discernment of bullshit from lies very difficult. Furthermore, both the bullshitter and liar can say the very same things and appear to be genuinely concerned with communicating the truth. It is important to note, however, that our participants were not privy to which statements, in particular, were false, lies, or bullshit—they were only made aware that some of the information they were exposed to during the interest-rating task was false, lies, or bullshit. Thus, it is unclear whether the effects demonstrated here extend to instances whereby people are not privy to the intentions of the bullshitter/liar; in fact, we see no readily obvious reason to expect them to. That is, it seems that a person's awareness of the intent of the communicator (i.e., bullshitting or lying) is required for the illusory truth effects, as we have reported here, to emerge.

Limits on generality

While our findings support Frankfurt's (1986) theoretical position on the insidiousness of bullshit relative to lying, we acknowledge that individual differences in cognitive style, motivation, and contextual factors may moderate these effects. Future research should examine when and for whom bullshit is most insidious, with particular attention to variables such as epistemic motivation, domain knowledge, and the perceived stakes of the communicative context.

A related limitation concerns the operationalization of bullshit in the present experiments. Because the manipulation relied on how the source was described rather than on actual differences in the statements themselves, the present

findings speak to the effects of being labeled a bullshitter rather than to the effects of bullshit communication per se. Participants may have interpreted the label “bullshitter” in varied ways—for instance, as signaling incompetence, exaggeration, or general dishonesty—rather than in the precise Frankfurterian sense of truth-indifference. Although the additional clarification provided in the bullshitter condition (i.e., that bullshitters communicate without regard for truth or evidence) was intended to standardize this interpretation, individual variation in how the label is understood may have introduced noise. Future research would benefit from designs in which bullshit is behaviorally instantiated—for example, by having participants interact with sources whose communications are demonstrably truth-indifferent—rather than relying solely on source labels, thereby strengthening the causal link between truth-indifferent communication and its effects on perceived truth.

Our findings regarding the role of initial exposure conditions also align with the moderator analyses reported by Ye et al. (2026), whose meta-analysis of 182 studies found that veracity cues and longer presentation time during first exposure were among the strongest reducers of the illusory truth effect. Their results underscore the practical importance of what people encounter at the *moment of first exposure*—a window of opportunity that, once closed, becomes difficult to reopen through after-the-fact corrections. Relatedly, Lacassagne et al. (2022) demonstrated that even highly implausible statements are not immune to repetition effects when exposure is sufficiently frequent and measurement is sufficiently sensitive. Together, these findings suggest that the harmful influence of bullshit may extend even to claims that initially seem too outlandish to be believed, provided the conditions of repeated exposure are met.

A further limitation of the present research is our use of primarily neutral, factual statements as experimental stimuli. While this approach provided good experimental control, real-world bullshit often involves controversial or emotionally charged topics, from political ideology to public health issues. Future research would do well to examine whether the enhanced illusory truth effect we observed for bullshit versus lies extends to more contentious domains where motivated reasoning and emotional investment may influence truth judgments. For instance, the relative impact of bullshitting versus lying might differ when addressing hot-button political issues, conspiracy theories, or health misinformation. Such research could help determine whether emotional engagement with the content moderates the insidious effects of bullshit we observed here.

Finally, our findings from Experiments 2 and 3 suggest a potential avenue for countering the harmful effects of bullshit: forewarning people about the nature of communication they will encounter. When participants were

informed about the presence of bullshit or lies before initial exposure, the illusory truth effect was attenuated. This observation points to the possibility that educational interventions teaching people to recognize bullshit, combined with clear labeling of potentially deceptive content, might help reduce bullshit’s influence on belief formation. However, given that people rarely receive such explicit warnings in real-world contexts, future research should explore how to effectively implement such preventive strategies without compromising authentic communication.

Conclusion

These results of three studies suggest the conceptual distinction between lying and bullshitting is not meaningless semantics, but a meaningful difference with important consequences for all forms of communication. Our investigation showed through the application of the illusory truth procedural paradigm that not only is it easier to dismiss lies compared to bullshit, but that bullshit is perceived as more truthful than lies—repeated exposure to bullshit increased its perceived truthfulness compared to repeated exposure to lies when cues to deception come after initial exposure. Support for both the lie dismissal and insidious bullshit hypotheses, suggest that being exposed to bullshit will have more deleterious effects on judgment and decision making relative to lies. Thus, compared to lying, bullshitting can be a more insidious form of communication by distorting the very beliefs upon which people base their judgments and decisions—and what people believe to be true is foundational to optimal decision making (Alcock, 2018). In a world where truth-indifferent communication is pervasive and often goes unrecognized, these findings carry a sobering message: the communicators least concerned with truth may exert the greatest influence on what the rest of us come to believe. Understanding, detecting, and countering the insidious effects of bullshit is not merely an academic exercise—it is an increasingly urgent challenge for everyday reasoning, public discourse, and the health of democratic deliberation.

Author Contributions John V. Petrocelli: Conceptualization, Methodology, Formal Analysis, Writing – Original Draft, Writing – Review & Editing, Supervision, Project Administration.

Elijah N. Rice: Data Curation, Formal Analysis, Writing – Review & Editing.

Joseph M. Curran: Data Collection, Writing – Review & Editing.

Funding Open access funding provided by the Carolinas Consortium. The authors received no financial support for the research, authorship, and/or publication of this article.

Data availability All data and procedures are available online at: <https://osf.io/3syxn/overview>.

Declarations

Ethical Standards This research involved Human Participants and was approved by the Human Subjects Committee of the Institutional Review Board of Wake Forest University.

Consent to Publish Not applicable. This study did not include any individual person's data in any form (e.g., individual details, images, or videos) that would require explicit consent to publish.

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Albarracín, D., Kumkale, G. T., & Vento, P. P. D. (2017). How people can become persuaded by weak messages presented by credible communicators: Not all sleeper effects are created equal. *Journal of Experimental Social Psychology*, 68, 171–180.
- Alcock, J. E. (2018). *Belief: What it means to believe and why our convictions are so compelling*. Prometheus Books.
- Arico, A. J., & Fallis, D. (2013). Lies, damned lies, and statistics: An empirical investigation of the concept of lying. *Philosophical Psychology*, 26, 790–816.
- Arkes, H. R., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, 2, 81–94.
- Ashton, M. C., Lee, K., & Visser, B. A. (2023). Does higher H mean less BS? Relations of a misleading communication style with the HEXACO personality factors. *Current Psychology*, 42, 23040–23044.
- Augustine, A. (1887). On lying. In P. Schaff & K. Knight (Eds.), *Nicene and Post-Nicene fathers, first series* (Vol. 3) (pp. 457–477). Christian Literature Publishing.
- Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, 5, 241–252.
- Begg, I., Anas, A., & Farinacci, S. (1992). Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth. *Journal of Experimental Psychology: General*, 121, 446–458.
- Blötnner, C., & Bergold, S. (2023). It is double pleasure to deceive the deceiver: Machiavellianism is associated with producing but not necessarily with falling for bullshit. *British Journal of Social Psychology*, 62, 467–485.
- Bond, C. F. Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10, 214–234.
- Bond, C. F. Jr., & DePaulo, B. M. (2008). Individual differences in judging deception: Accuracy and bias. *Psychological Bulletin*, 134, 477–492.
- Brashier, N. M., Umanath, S., Cabeza, R., & Marsh, E. J. (2017). Competing cues: Older adults rely on knowledge in the face of fluency. *Psychology and Aging*, 32, 331–337.
- Castelli, P., & Ghetti, S. (2014). Resisting imagination and confabulation: Effects of metacognitive training. *Journal of Experimental Child Psychology*, 126, 339–356.
- Cook, T. D., Gruder, C. L., Hennigan, K. M., & Flay, B. R. (1979). History of the sleeper effect: Some logical pitfalls in accepting the null hypothesis. *Psychological Bulletin*, 86, 662–679.
- De keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2020). Investigating the robustness of the illusory truth effect across individual differences in cognitive ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin*, 46, 204–215.
- DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology*, 70, 979–995.
- DePaulo, B. M., Charlton, K., Cooper, H., Lindsay, J. J., & Muhlenbruck, L. (1997). The accuracy–confidence correlation in the detection of deception. *Personality and Social Psychology Review*, 1, 346–357.
- Ekman, P. (1985). *Telling lies: Clues to deceit in the marketplace, marriage, and politics*. W. W. Norton and Company.
- Ekman, P. (2009). *Telling lies: Clues to deceit in the marketplace, politics, and marriage*. W.W. Norton & Company.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191.
- Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science*, 21, 801–803.
- Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144, 993–1002.
- Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin and Review*, 26, 1705–1710.
- Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3, 552–564.
- Fisher, M., & Keil, F. C. (2016). The curse of expertise: When more knowledge leads to miscalibrated explanatory insight. *Cognitive Science*, 40, 1251–1269.
- Foos, A. E., Keeling, K., & Keeling, D. (2016). Redressing the sleeper effect: Evidence for the favorable persuasive impact of discounting information over time in a contemporary advertising context. *Journal of Advertising*, 45, 19–25.
- Frankfurt, H. (1986). On bullshit. *Raritan Quarterly Review*, 6, 81–100.
- Gilbert, D. T. (1991). How mental systems believe. *American Psychologist*, 46, 107–119.
- Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of Personality and Social Psychology*, 59, 601–613.
- Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (Vol. 3, pp. 41–58). Seminar.
- Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16, 107–112.

- Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, 6, Article 38.
- Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, 19, 212–225.
- Henderson, E. L., Simons, D. J., & Barr, D. J. (2021). The trajectory of truth: A longitudinal study of the illusory truth effect. *Journal of Cognition*, 4, Article 29.
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15, 635–650.
- Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, 3, 306–340.
- Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin*, 13, 314–336.
- Jacoby, L. L., Kelley, C. M., & Dywan, J. (1989). Memory attributions. In H. L. Roediger, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 391–422). Erlbaum.
- Jarvis, W. B. G. (2016). *MediaLab v2016*. Empirisoft.
- Johar, G. V., & Roggeveen, A. L. (2007). Changing false beliefs from repeated advertising: The role of claim-refutation alignment. *Journal of Consumer Psychology*, 17, 118–127.
- Kimbrough, S. (2006). On letting it slide. In G. L. Hardcastle, & G. A. Reisch (Eds.), *Bullshit and philosophy: Guaranteed to get perfect results every time* (pp. 3–18). Open Court.
- Kominsky, J. F., & Keil, F. C. (2014). Overestimation of knowledge about word meanings: The misplaced meaning effect. *Cognitive Science*, 38, 1604–1633.
- Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 107–118.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77, 1121–1134.
- Lacassagne, D., Béna, J., & Corneille, O. (2022). Is Earth a perfect square? Repetition increases the perceived truth of highly implausible statements. *Cognition*, 223, Article 105052.
- Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology*, 32, 951–955.
- Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory and Cognition*, 34, 1667–1675.
- Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., & Zittrain, J. L. (2018). The science of fake news. *Science*, 359, 1094–1096.
- Lee, J., & Bissell, K. (2024). Correcting vaccine misinformation on social media: The inadvertent effects of repeating misinformation within such corrections on COVID-19 vaccine misperceptions. *Current Psychology*, 43(26), 22754–22766.
- Levine, T. R., & McCormack, S. A. (1992). Linking love and lies: A formal test of the McCormack and Parks model of deception detection. *Journal of Social and Personal Relationships*, 9, 143–154.
- Levine, T. R., Park, H. S., & McCormack, S. A. (1999). Accuracy in detecting truths and lies: Documenting the veracity effect. *Communication Monographs*, 66, 125–144.
- Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33, 378–392.
- Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13, 106–131.
- Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the post-truth era. *Journal of Applied Research in Memory and Cognition*, 6, 353–369.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.
- Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance*, 26, 149–171.
- Littrell, S., & Fugelsang, J. A. (2024). Bullshit blind spots: The roles of miscalibration and information processing in bullshit detection. *Thinking and Reasoning*, 30, 49–78.
- Littrell, S., Risko, E. F., & Fugelsang, J. A. (2021). The Bullshitting Frequency Scale: Development and psychometric properties. *British Journal of Social Psychology*, 60, 248–270.
- Mahon, J. E. (2008). Two definitions of lying. *International Journal of Applied Philosophy*, 22, 211–230.
- Martin, L. L. (1986). Set/reset: Use and disuse of concepts in impression formation. *Journal of Personality and Social Psychology*, 51, 493–504.
- Martin, L. L., Seta, J. J., & Crelia, R. A. (1990). Assimilation and contrast as a function of people's willingness and ability to expend effort in forming an impression. *Journal of Personality and Social Psychology*, 59, 27–37.
- Meibauer, G. (2021). Ambiguous specificity: The production of foreign policy bullshit in electoral contexts. *Politics*, 41, 15–30.
- Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29, 665–675.
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50.
- Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, 88, 185–200.
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, 10, 549–563.
- Petrocelli, J. V. (2018). Antecedents of bullshitting. *Journal of Experimental Social Psychology*, 76, 249–258.
- Petrocelli, J. V. (2021a). Bullshitting and persuasion: The persuasiveness of a disregard for the truth. *British Journal of Social Psychology*, 60, 1464–1483.
- Petrocelli, J. V. (2021b). *The life-changing science of detecting bullshit*. St. Martin's.
- Petrocelli, J. V., Watson, H. F., & Hirt, E. R. (2020). Self-regulatory aspects of bullshitting and bullshit detection. *Social Psychology*, 51, 239–253.
- Petrocelli, J. V., Seta, C. E., & Seta, J. J. (2023a). Lies and bullshit: The negative effects of misinformation grow stronger over time. *Applied Cognitive Psychology*, 37, 409–418.
- Petrocelli, J. V., Silverman, H. E., & Shang, S. X. (2023b). Social perception and influence of lies vs. bullshit: A test of the insidious bullshit hypothesis. *Current Psychology*, 42, 9609–9617.
- Pratkanis, A. R., Greenwald, A. G., Leippe, M. R., & Baumgardner, M. H. (1988). In search of reliable persuasion effects: III. The sleeper effect is dead: Long live the sleeper effect. *Journal of Personality and Social Psychology*, 54, 203–218.
- Priester, J., Wegener, D., Petty, R., & Fabrigar, L. (1999). Examining the psychological process underlying the sleeper effect: The elaboration likelihood model explanation. *Media Psychology*, 1, 27–48.

- Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition: An International Journal*, 8, 338–342.
- Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, 86, 61–79.
- Reisch, G. A. (2006). The pragmatics of bullshit, intelligently designed. In G. L. Hardcastle, & G. A. Reisch (Eds.), *Bullshit and philosophy: Guaranteed to get perfect results every time* (pp. 33–47). Open Court.
- Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review*, 31, 180–206.
- Rothbart, M., & Park, B. (1986). On the confirmability and disconfirmability of trait concepts. *Journal of Personality and Social Psychology*, 50, 131–142.
- Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science*, 26, 521–562.
- Rutschmann, R., & Wiegmann, A. (2017). No need for an intention to deceive? Challenging the traditional definition of lying. *Philosophical Psychology*, 30, 438–457.
- Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change*. Yale University Press.
- Sherif, M., & Sherif, C. W. (1967). Attitudes as the individual's own categories: The social-judgment approach to attitude and attitude change. In C. W. Sherif, & M. Sherif (Eds.), *Attitude, ego-involvement and change* (pp. 105–139). Wiley.
- Sherif, C. W., Sherif, M., & Nebergall, R. E. (1965). *Attitude and attitude change: The social judgment-involvement approach*. W. B. Saunders.
- Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G., & Wilson, D. (2010). Epistemic vigilance. *Mind and Language*, 25, 359–393.
- Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, 45, 1115–1143.
- Turri, A., & Turri, J. (2015). The truth about lying. *Cognition*, 138, 161–168.
- Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning Memory and Cognition*, 33, 219–230.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359, 1146–1151.
- Wang, W. C., Brashier, N. M., Wing, E. A., Marsh, E. J., & Cabeza, R. (2016). On known unknowns: Fluency and the neural mechanisms of illusory truth. *Journal of Cognitive Neuroscience*, 28, 739–746.
- Wegener, D. T., & Petty, R. E. (1995). Flexible correction processes in social judgment: The role of naive theories in corrections for perceived bias. *Journal of Personality and Social Psychology*, 68, 36–51.
- Williams, B. (2002). *Truth and truthfulness: An essay in genealogy*. Princeton University Press.
- Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, 116, 117–142.
- Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines*, 8, 137–159.
- Ye, S., Attali, D., Ghazi, M., Cachia, A., Cassotti, M., & Borst, G. (2026). Systematic review and meta-analysis of the evidence for an illusory truth effect and its determinants. *Nature Communications*, 17, Article3270.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.