UNIVERSITATEA BABEŞ-BOLYAI
BABEŞ-BOLYAI TUDOMÁNYEGYETEM
BABEŞ-BOLYAI UNIVERSITÄT
TRADITIO ET EXCELLENTIA

APPLIED COGNITIVE PSYCHOLOGY DOCTORAL SCHOOL

SUMMARY OF THE Ph.D. THESIS

COGNITIVE DYNAMICS IN FAKE NEWS VULNERABILITY:

REPETITION, REFLECTION, AND RECALIBRATION

AUTHOR: Ph.D. CANDIDATE SECARĂ EUGEN-CĂLIN

SCIENTIFIC ADVISOR: PROFESSOR NICOLAE-ADRIAN OPRE, Ph.D.

CLUJ-NAPOCA

2026

Contents

1. Introduction

The digital transformation of contemporary information environments has fundamentally altered how individuals encounter, evaluate, and act upon information. Unlike traditional media ecosystems, where editorial gatekeeping imposed at least partial control over information quality and dissemination, contemporary digital environments allow professionally produced journalism, opinion pieces, personal testimony, satire, advertising, algorithmically amplified content, and deliberately fabricated narratives to coexist within the same informational space (Spohr, 2017; Nguyen, 2020). For information consumers, these boundaries are often blurred, increasing the cognitive demands associated with evaluating credibility, trustworthiness, and relevance.

Exposure to false or misleading information has been associated with political polarization, erosion of institutional trust, conspiracy ideation, distorted health beliefs, and maladaptive health behaviors (Pennycook & Rand, 2019; Loomba et al., 2021; Allington et al., 2021; Lee et al., 2020). While misinformation itself is not historically novel, the scale, speed, and persistence with which misleading content circulates in digitally mediated ecosystems have transformed it into a major societal concern. The COVID-19 pandemic provided a particularly salient example of this phenomenon, generating what the World Health Organization described as an "infodemic," characterized by uncertainty, conflicting claims, emotionally charged messaging, and widespread health misinformation that directly influenced public behavior (Allington et al., 2021; Naveed et al., 2021).

Within this context, understanding why some individuals are more vulnerable to fake news than others has become an urgent research priority. Existing work has made important progress in identifying cognitive mechanisms associated with fake news vulnerability. A substantial body of evidence suggests that individuals who engage more readily in reflective, analytic reasoning are generally better able to distinguish between true and false information,

both in political and health-related domains (Pennycook & Rand, 2018, 2019, 2020; Scherer et al., 2020). These findings have strongly influenced contemporary theoretical models, many of which conceptualize fake news vulnerability primarily as a consequence of insufficient analytic engagement.

The qualities of the online environment also contribute to the issue. People often encounter information repeatedly, in fragmented formats, under conditions of distraction, emotional activation, time pressure, and social reinforcement (Pennycook et al., 2018, 2021). Under such conditions, investment of trust is also influenced by familiarity, processing fluency, narrative coherence, pre-existing beliefs, and face validity (i.e., intuitive plausibility, Hasher et al., 1977; Unkelbach & Greifeneder, 2013; Fazio et al., 2015, 2019; Pennycook et al., 2018). These influences have been hypothesized to be further amplified by algorithmically curated digital environments that selectively reinforce exposure patterns based on engagement metrics rather than informational quality (Spohr, 2017; Nguyen, 2020).

Beyond cognitive style, individual differences in interpretative tendencies also appear relevant. Research has linked fake news vulnerability to apophenic tendencies, including ontological confusion and receptivity to pseudo-profound bullshit, suggesting that susceptibility may reflect broader cognitive styles characterized by reduced sensitivity to semantic, causal, or ontological inconsistencies (Bronstein et al., 2019; Pennycook et al., 2015). Similarly, prior work has suggested that metacognitive factors, particularly overconfidence in one's own judgments, may play an important role in fake news vulnerability (De Keersmaecker & Roets, 2017; Lyons et al., 2021; Vranic et al., 2022). Yet these processes remain comparatively underexplored relative to analytic reasoning.

A further limitation of the existing literature concerns ecological validity. Much experimental misinformation research relies on simplified headline-based paradigms, often presented in

decontextualized formats that differ substantially from real-world information encounters (Pennycook et al., 2018, 2021). While these paradigms offer strong experimental control, they necessarily constrain the range of cognitive processes that can be observed. Full-length misleading content allows for conflict detection across text segments, narrative integration, and metacognitive monitoring during reading in ways that headlines alone cannot capture.

Similarly, although repeated exposure is central to contemporary digital misinformation environments, its effects are often modeled experimentally over little time or few repetitions. The illusory truth effect demonstrates that repeated exposure increases perceived truthfulness through familiarity and fluency mechanisms (Hasher et al., 1977; Fazio et al., 2019), while related work on the continued influence effect suggests that misinformation can shape reasoning even after correction (Johnson & Seifert, 1994; Walter & Murphy, 2018; O'Rear & Radvansky, 2020). Yet relatively few studies have examined how repeated exposure to ecologically realistic misinformation shapes attitudes longitudinally, particularly in health-related domains.

Health misinformation represents a particularly important domain for investigation. Unlike many forms of political misinformation, health-related falsehoods can directly influence treatment decisions, adherence to medical recommendations, vaccination behavior, and preventive health actions (Loomba et al., 2021; Allington et al., 2021). At the same time, health-related fake news frequently intersects with complementary and alternative medicine narratives, distrust in conventional medicine, and conspiracy-oriented explanatory frameworks, making it potentially theoretically distinct from other misinformation domains both on the level of influences and on the level of practical implications(Scherer et al., 2021).

Taken together, the literature reveals several important gaps. Existing models have disproportionately emphasized analytic reasoning, while comparatively neglecting memory

processes, metacognitive monitoring, and environmental exposure dynamics. Ecological validity remains limited due to reliance on simplified stimuli, repeated exposure effects remain insufficiently modeled, and relatively little work integrates fake news vulnerability with real-world health beliefs and behaviors.

The present thesis addresses these limitations through a multi-study empirical investigation of health-related fake news vulnerability. It advances an integrative framework in which fake news vulnerability emerges from interactions among cognitive style, memory processes, metacognitive monitoring, personality-linked interpretative tendencies, domain-relevant knowledge (e.g., news media literacy and medical knowledge), and the structural properties of digital information environments. Across three complementary studies, the thesis examines how individuals evaluate full-length misleading content, provides empirical tests of influential conceptual frameworks regarding fake news, investigates whether overconfidence represents a distinct and mutable vulnerability pathway, and explores how repeated exposure shapes attitudes over time under ecologically relevant conditions.

By integrating cognitive, metacognitive, and environmental perspectives within a unified empirical program, the thesis aims to provide a more comprehensive understanding of why fake news is compelling, persistent, and difficult to correct.

2. Theoretical background and knowledge gaps

The study of fake news vulnerability has expanded rapidly in recent years, driven by growing societal concerns regarding political polarization, public mistrust, conspiracy ideation, and the health consequences of misleading information (Pennycook & Rand, 2019; Loomba et al., 2021; Allington et al., 2021). A central question in this literature concerns why some individuals are more likely than others to trust false or misleading information. One of the most influential explanations emphasizes differences in reasoning style. According to analytic reasoning accounts, susceptibility to misinformation reflects insufficient engagement

in reflective, effortful cognitive processing, leading individuals to rely excessively on intuition and fast judgements when evaluating truth claims (Pennycook & Rand, 2018, 2019, 2020). Consistent with this perspective, cognitive reflection has repeatedly emerged as a robust predictor of truth discernment across a range of misinformation paradigms (Scherer et al., 2020).

These findings align with broader dual-process frameworks, which distinguish between fast, intuitive, heuristic processing and slower, more deliberative analytic reasoning (Kahneman, 2011; Evans, 2008; Evans and Stanovich, 2013). From this perspective, fake news vulnerability emerges when intuitive judgments are accepted without sufficient reflective scrutiny. Closely related conflict-detection accounts suggest that analytic reasoning is not necessarily absent in vulnerable individuals, but may fail to engage when misleading information does not trigger sufficient signals of inconsistency or uncertainty (De Neys, 2014).

However, the importance of context must not be underestimated, both historical (i.e., contexts encountered during the individual's learning history) and current. Modern news environments exposed to information in fragmented formats, optimized for fast, superficial processing, providing ample social cues and using language meant to impress rather than inform (Pennycook et al., 2018, 2021; Fazio et al., 2019; Spohr, 2017; Nguyen, 2020). The influence of historic contexts is reflected through beliefs, familiarity, and confidence in one's own world model. One important line of work referring to the importance of history has examined apophenic tendencies, defined as the propensity to detect meaningful patterns in ambiguous or even random information (Bainbridge et al., 2018). Closely related constructs, such as ontological confusion, involving inappropriate attribution of agency or intentionality, and receptivity to pseudo-profound bullshit (i.e., considering randomly generated quotes to be profound), have been linked to greater susceptibility to misinformation and conspiratorial

beliefs (Pennycook et al., 2015; Bronstein et al., 2019; Lobato et al., 2014; Van den Bulck & Custers, 2010). These findings suggest that vulnerability may partly reflect stable interpretative styles characterized by reduced sensitivity to semantic, causal, or ontological inconsistencies.

Health misinformation frequently relies on emotionally compelling narratives, anecdotal evidence, distrust in conventional medicine, and explanatory frameworks resembling those found in complementary and alternative medicine discourse (Scherer et al., 2021). Such content often appeals not only to intuitive plausibility, but also to broader worldviews concerning medicine, authority, and hidden forces. This suggests that health-related fake news vulnerability may be particularly suited to study the interactions between cognitive ability, domain-specific beliefs, and personality variables.

Another important but comparatively understudied component concerns knowledge structures. News media literacy has often been conceptualized as a protective factor against misinformation, reflecting knowledge about journalistic standards, sourcing practices, and media production processes. However, its role may be more nuanced than a simple protective factor. News media literacy may operate as a contextual scaffold that increases the probability of conflict detection by providing structural cues that something is implausible, manipulative, or inconsistent with normative news practices. Similarly, domain-specific knowledge, such as medical understanding, may shape how individuals interpret health claims, detect inconsistencies, or defer to seemingly authoritative but misleading narratives.

Repetition strongly influences credibility judgments beyond what can be explained through analytical reasoning and cognitive ability, by using fundamental mechanisms such as processing fluency (Unkelbach & Greifeneder, 2013). The effect of repetition on belief has been termed the illusory truth effect (Hasher et al., 1977; Fazio et al., 2015, 2019; Pennycook

et al., 2018). Closely related mere exposure effects demonstrate that repeated contact with stimuli can alter attitudes even in the absence of substantive persuasion (Zajonc, 1968; Bornstein & Craver-Lemley, 2022).

A related but distinct dimension concerns metacognitive monitoring: the ability to accurately assess the reliability of one's own cognitive processes (Fleming & Lau, 2014). Overconfidence has emerged as a promising explanatory factor in misinformation research, with prior work suggesting that individuals who overestimate their reasoning accuracy may be more susceptible to false claims (De Keersmaecker & Roets, 2017; Lyons et al., 2021; Vranic et al., 2022). This line of work is particularly relevant given the growing literature on accuracy nudges, which has shown that directing attention toward accuracy can reduce misinformation sharing and improve discernment both in experimental settings and ecologically valid digital environments (Pennycook et al., 2021; Mirhoseini et al., 2023). At the same time, explicit memory processes have independently been implicated in misinformation vulnerability through failures of belief updating and the persistence of misinformation even after correction, as demonstrated by the Continued Influence Effect literature (Johnson & Seifert, 1994; Walter & Murphy, 2018; O'Rear & Radvansky, 2020). These converging lines of research suggest that metacognitive distortions at the level of memory may represent a theoretically important but underexplored predictor of fake news vulnerability. Evaluating misinformation requires maintaining coherence across text, comparing newly encountered claims against previously stored knowledge, and detecting uncertainty when inconsistencies arise. From a conflict-detection perspective (De Neys, 2014), overconfidence in working memory may reduce sensitivity to local inconsistencies within a text, decreasing the likelihood of rereading or reassessing earlier information. Similarly, overconfidence in long-term memory may reduce the probability of questioning claims that conflict with prior knowledge, limiting verification or reflective scrutiny. In this

framework, misleading information may be accepted not because reasoning capacity is absent, but because overconfidence suppresses the epistemic uncertainty necessary to initiate deeper evaluation.

Taken together, the literature suggests that fake news vulnerability is a multifaceted phenomenon shaped by interacting cognitive, metacognitive, and environmental influences. While substantial progress has been made in identifying the role of analytic reasoning and related cognitive processes, several theoretically relevant areas remain comparatively less explored, particularly the contribution of memory-specific metacognitive monitoring, ecologically richer assessment paradigms, and the cumulative effects of repeated exposure in realistic informational environments. In addition, although health misinformation has attracted increasing scholarly attention due to its direct societal relevance, fewer studies have integrated cognitive vulnerability with consequential real-world attitudes and behaviors. The present thesis builds on these existing lines of research through an integrative empirical program combining cognitive, metacognitive, ecological, and behavioral perspectives.

To the best of our knowledge, fake news vulnerability has been assessed almost exclusively through variations of a single paradigm, initially introduced in Pennycook et al. (2018) and systematically reviewed in Pennycook et al. (2021). Motivated by the finding that approximately 59 percent of links shared on Twitter were not opened before being reposted (Gabielkov et al., 2016), the paradigm presents participants with news headlines without granting access to the full articles and asks participants about their reliability or accuracy. This format has been used in its most parsimonious form by Clayton et al. (2019), who presented only the headlines.

Several studies expand this basic design with additional elements such as images and short descriptions (mirroring a Facebook-style post; Pennycook et al., 2018) or source attributions

(Pennycook and Rand, 2019). Headlines are typically drawn from fact-checking databases, and Pennycook et al. (2021) emphasize that the validity of the entire paradigm depends on selecting ecologically relevant stimuli. They recommend choosing headlines that are familiar enough for participants to recognize their genre and plausibility, while avoiding overexposure, especially in the case of true headlines.

The Facebook-like format further enhances ecological validity, given that social media platforms constitute the environments in which fake news is most frequently encountered. Pretesting stimuli is strongly recommended, especially when constructing category-specific sets (for example, politically liberal versus conservative fake news). The number of items used influences measurement precision, although overly long item sets may introduce fatigue and disengagement.

Using Signal Detection Theory (Green and Swets, 1966), studies usually compute two indices that reflect overall sensitivity to truth versus falsehood and response bias (Batailler et al., 2021; Pennycook et al., 2021). Research using this paradigm has provided valuable insight into the cognitive and motivational mechanisms underlying fake news vulnerability, which will be reviewed in the following sections.

To the best of our knowledge, only four studies to date have presented participants with full-length fake news articles. Porter et al. (2018), who examined the effectiveness of corrections on political misinformation. Participants were shown two of six fabricated political stories, either with or without a correction, and presented either in text or video format. Their findings were broadly consistent with those obtained from the headline paradigm, suggesting some degree of convergent validity. Another study presenting full disinformation articles is that of Schaewitz et al. (2020), who experimentally manipulated message characteristics within two policy fabricated full-length articles (i.e., crime policy, elderly care policy),

including source credibility, sensationalism, subjectivity, inconsistencies, and image manipulation. Participants rated each article's credibility, accuracy, and likelihood of sharing. Their findings showed that message cues had comparatively small effects, whereas individual differences, especially need for cognition and topic-related attitudes, were stronger predictors of perceived credibility and sharing intentions. Pehlivanoglu et al. (2021) examined how analytical reasoning and manipulated source credibility influence the evaluation of 12 full-length real and fake news stories. The stories covered politically charged topics such as claims about the Black Lives Matter movement or gun-confiscation policies, health-related narratives including allegations that doctors refused medical care on religious grounds, and religiously themed stories involving Mormonism, same-sex marriage, or statements falsely attributed to Pope Francis. Participants read each full article and rated its accuracy and credibility. Across two independent studies, higher analytical reasoning consistently predicted better discrimination of fake from real articles, while source credibility also shaped judgments, particularly among participants lower in analytic reasoning. Finally, Torreggiani (2025) developed a dataset of 80 authentic social media posts drawn from U.S. and Italian political and news-related events (e.g., sports, celebrities). Participants in both countries were shown a randomized subset of these posts and asked to judge their accuracy, quality, and perceived political leaning. Findings mostly aligned with prior findings: participants distinguished real from fake content reasonably well, cognitive style predicted accuracy judgments, however, politically congruent misinformation showed only weak effects.

The headline-based approach offers important advantages in terms of feasibility and participant burden, and its results are generally aligned with theoretical expectations (Pennycook et al., 2021). However, full articles provide additional cues that can trigger conflict detection, enhance analytic processing, and support more ecologically complex reasoning. One example is provided by Dogo et al. (2020), who using topic-modelling across

seven cross-domain datasets showed that the opening sentences of fake articles tend to be vaguer, less consistent, and more weakly connected to the remainder of the text, whereas real articles display stronger narrative continuity and coherence. These internal structural properties might serve as important triggers for analytic processing and cannot be detected through headline-only measures. For these reasons, more comprehensive assessment paradigms that incorporate full articles should be considered in replications and in future research.

Political content remains the dominant topic in fake news research. This poses two problems. First, political misinformation is strongly tied to local context, making cross-cultural comparisons difficult because new stimulus sets must be developed for each setting, introducing error and limiting comparability. Second, political fake news stories tend to be short-lived, as political actors and issues change rapidly, meaning that headline sets must be continuously revalidated.

By contrast, health-related fake news exhibits greater temporal stability. Once introduced, health misinformation often resurfaces during crises or waves of renewed concern and may persist even after thorough debunking (for example, anti-vaccination claims; Rao and Andrade, 2011). This persistence makes health misinformation a valuable and underused domain for examining cognitive vulnerability to fake news.

Given these limitations, a more diversified approach to measuring fake news vulnerability is warranted. Broadening stimulus domains beyond politics, incorporating full articles, and developing cross-contextual and longitudinal assessment tools will increase ecological validity and strengthen the theoretical inferences that can be drawn regarding why people fall for misinformation.

3. Conceptual framework and research objectives

Given the current state of the literature, including recent evidence on the mixed effectiveness, contextual dependencies, and potential spillover effects of existing interventions, the main objective of the present project is to identify and analyze additional cognitive mechanisms and environmental factors that contribute to individual vulnerability to fake news, and to test the efficiency of interventions that target these processes. To integrate the multitude of theoretical perspectives and empirical findings reviewed in this thesis, and to clarify the logic underlying the empirical studies presented, a conceptual model is proposed in Figure 1. The model represents the cognitive, metacognitive, personality-related, and environmental factors that jointly contribute to fake news vulnerability, as well as the pathways through which these factors are hypothesised to interact. It serves as a guiding framework that delineates the main domains of influence and their interrelations. In doing so, the model provides both a synthesis of existing theory and a structured roadmap for the original research contributions developed across the three studies.

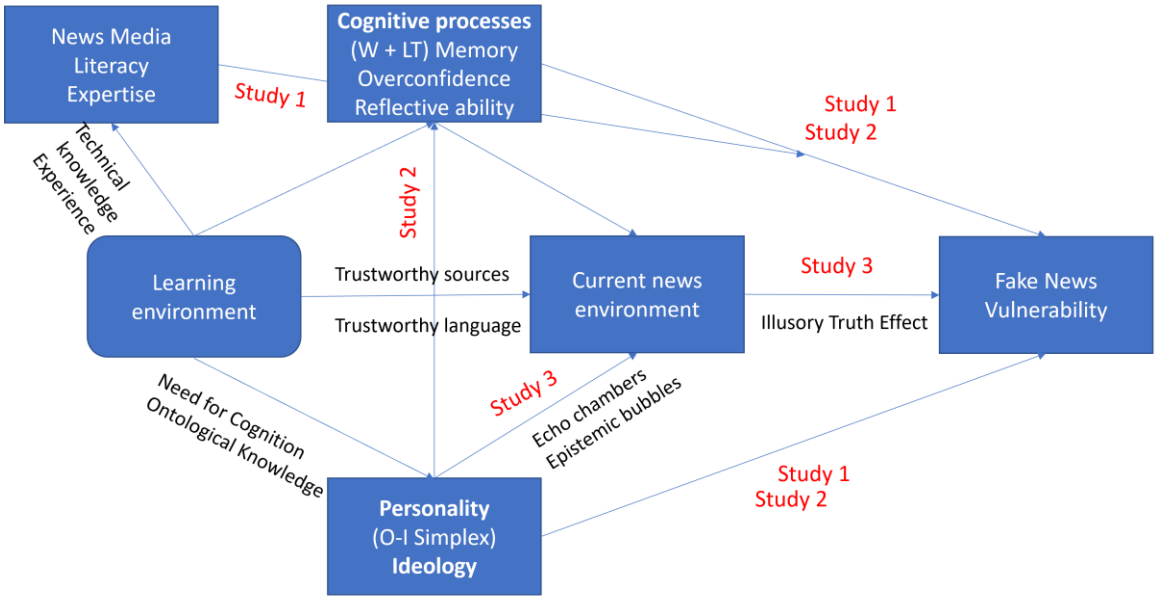


Figure 1. Conceptual model of the thesis

As the issue of fake news vulnerability cannot be effectively addressed by focusing solely on individual characteristics while ignoring environmental influences, to the proposed model

broadens the scope of analysis to include the effects of individuals' learning histories. Distal environmental experiences shape four key domains that influence susceptibility to misinformation.

First, life history shapes current news environments by teaching individuals which sources are trustworthy and which structural cues signal the need for more rigorous analysis, such as linguistic features, narrative style, or article structure.

Second, learning histories contribute to the development of personality and ideology, including tendencies toward deep or shallow cognitive processing, such as need for cognition (Cacioppo and Petty, 1982), and the formation of world-models against which the validity of proximal information can be assessed (e.g., ontological expectations about which agents can perform particular actions).

Third, prior exposure and experience shape domain expertise and news media literacy, including knowledge of subject-matter content and the conventions of news production.

Fourth, life history influences metacognitive ability, including the capacity to mobilize appropriate cognitive processes and to evaluate the success of one's reasoning. Reinforcement histories associated with analytic thinking and heuristic use shape how often such abilities are practiced and refined.

One consistent finding in the literature is that repeated exposure to information increases perceived accuracy, even when the information is known to be false (Brown and Nix, 1996; Fazio et al., 2019; Fazio et al., 2020). This *illusory truth effect* (see section 1.3.2.) also extends to fake news, with as few as two prior exposures being sufficient to boost believability (Pennycook et al., 2018). Crucially, the effect appears independent of knowledge, cognitive ability, or cognitive style (De Keersmaecker et al., 2020; Fazio et al.,

2015), which is why in the proposed conceptual model it is represented as a direct pathway from the current news environment to fake news vulnerability.

Digital environments in which misinformation spreads are further characterized by partial information blindness (Haim et al., 2018). This occurs through epistemic bubbles, which limit exposure to diverse viewpoints through personal or algorithmic filtering (Nguyen, 2020), and echo chambers, which go beyond selective exposure by actively discrediting out-group perspectives and granting high trust to in-group sources (Nguyen, 2020). Both structures reduce opportunities to encounter corrective information and trigger the illusory truth effect by providing repeated, unchallenged exposure to congruent content (Spohr, 2017). Adherence to epistemic bubbles and echo chambers is typically driven by shared beliefs, similar reasoning patterns, and common informational sources, closely aligning with personality and ideological factors frequently studied in relation to misinformation.

To date, the largest body of empirical research on fake news vulnerability has examined its relationship with cognitive processes (often through dual-processing models; Bronstein et al., 2020, Torreggiani, 2025), personality traits (particularly the Intellect-Openness simplex; DeYoung et al., 2012), and ideology (including debates on motivated System 2 reasoning; Tappin et al., 2020, Torreggiani, 2025). These findings will be reviewed in detail in subsequent sections. However, it is important to note that most of this evidence is correlational, and clear causal pathways have not yet been established.

The interaction between news media literacy and cognitive processing remains underexplored, as no study has empirically tested whether media literacy moderates the relationship between reflective ability and fake news vulnerability. From a theoretical standpoint, such moderation is plausible: individuals with limited reflective ability may rely more heavily on contextual cues derived from news-writing conventions, and increased

media literacy could compensate for lower analytic engagement. Conversely, as reflective abilities increase, the marginal contribution of media literacy may diminish, because analytic processing is more reliably triggered by the content itself.

Before examining the specific components and hypothesized interactions in the model, it is essential to consider how fake news vulnerability is measured, as valid and reliable measurement is a prerequisite for drawing meaningful conclusions about conceptual relationships or intervention effectiveness.

The three empirical studies presented in this thesis were conceived as a coordinated research program addressing multiple types of objectives. Each study targets specific theoretical, methodological, and practical questions related to fake news vulnerability, while collectively contributing to a coherent explanatory framework.

Objective 1. Theoretical objective

Identify cognitive mechanisms that predict fake news vulnerability beyond analytic reasoning.

Research on fake news has frequently relied on dual process accounts and has treated analytic reasoning as the central protective factor. However, evidence reviewed in Chapter I indicates that memory fluency, the state of the current news environment and metacognitive bias also play central roles.

Study 1 evaluates a comprehensive model of health-related fake news vulnerability that integrates cognitive reflection, personality factors from the Openness-Intellect simplex and news media literacy.

Study 2 evaluates the contribution of metacognitive monitoring by examining memory overconfidence and its relation to fake news vulnerability.

Study 3 examines how repeated exposure to polarized articles shapes explicit and implicit evaluations and contributes to fluency-based changes.

Objective 2. Methodological objective

Increase the ecological validity of fake news vulnerability assessments.

Traditional headline-based tasks provide valuable information but capture only a limited part of misinformation processing. They do not model temporal exposure, narrative coherence, memory updating or the implicit effects of repeated contact with complex articles.

Study 1 examines empirical evidence for the Information Disorder Framework through two specific objectives. First, it assesses whether misinformation and disinformation are associated with distinct patterns of vulnerability by analysing the factorial structure of full-length health-related fake news articles. Second, it evaluates the convergent validity of these measures by examining their associations with theoretically relevant cognitive, personality-related, and news media literacy constructs.

Study 2 employs the articles validated in the first study to assess cognitive phenomena that cannot be adequately captured through less information rich formats.

Study 3 moves beyond headline or single sentence judgments by using repeated exposure to full length articles over a two-week period, approximating algorithm driven information environments.

Objective 3. Practical objective

Assess interventions that reduce fake news vulnerability based on the identified cognitive and contextual factors.

Interventions that aim to reduce fake news vulnerability require clear targets. As reviewed in Chapter I, fact checking and warnings often have limited effects, while memory processes,

news media literacy and metacognitive calibration (including accuracy prompts) may be more promising intervention points.

Study 2 assesses the potential of overconfidence correcting feedback as an intervention mechanism and as an accuracy prompt.

Study 3 evaluates the protective role of a diverse informational diet on the effect of repeated exposure on attitude change.

4. General Methodological Approach

The original studies included in the thesis use complementary methodologies that examine fake news processing at different levels of analysis and across different ecological conditions. Across all studies, attention is given to increasing ecological validity while maintaining experimental control.

The thesis employs both exploratory and confirmatory factor analyses. Study 1 tests the factors underlying misinformation and disinformation and assesses the predictors of health-related fake news using an observational, cross-sectional design and a repeated measures design.

The project incorporates performance based metacognitive measures, an approach rarely applied in misinformation research. Study 2 examines confidence accuracy relations in working memory and long-term knowledge tasks to quantify metacognitive sensitivity and metacognitive bias. This allows the identification of calibration failures that may prevent individuals from detecting conflicts or revising misinformation. By using objective indices rather than self-report measures, the study increases the validity and interpretability of metacognitive constructs in relation to fake news. The use of an experimental pre-post design allows both the investigation of overconfidence as a predictor of fake news vulnerability (at

pre-test) and the evaluation of how effective overconfidence correcting feedback is in reducing fake news vulnerability (relative to a control intervention, at post-test).

The research also integrates controlled experimental exposure with repeated measures. Study 3 uses full length articles presented over a two-week period, allowing the examination of how repetition and fluency shape explicit and implicit attitudes. This approach captures processes that develop gradually through repeated contact with polarized content, a central feature of contemporary digital environments.

Finally, the project follows transparent and replicable research practices. All studies use pretested or well validated materials, establish measurement reliability, apply appropriate statistical controls and report effect sizes and confidence intervals. Studies 2 and 3 were preregistered. Together, these methodological principles create a coherent empirical foundation for the theoretical framework proposed in this thesis.

5. Original research contributions

Study 1. Exploring the Information Disorder Framework: Health-Related Fake News, Bullshit, and Their Impact on COVID-19 Protective Measures and Vaccination Intent

In this study, we investigated differences in responses to health-related misinformation and disinformation as defined by the Information Disorder framework (Wardle and Derakhshan, 2017). Across three substudies, we adapted the methodology used to study bullshit receptivity to create health-related bullshit and compared its perception to fake news articles, following guidelines for behavioral fake news research (Pennycook et al., 2021).

The first sub-study employed exploratory factor analysis (EFA) to investigate the latent structure underlying responses to health-related misinformation and disinformation using a repeated-measures design. Repeated exposure was used to capture stable patterns of vulnerability while accounting for within-person variability related to memory, familiarity,

and fluency effects. In addition to article-based responses, the design included theoretically related cognitive and personality measures to explore convergent validity. The primary hypothesis guiding this sub-study was that vulnerability to misinformation and disinformation would load onto separate factors, according to the conceptual distinction within the Information Disorder Framework. Secondary hypotheses were that vulnerability to misinformation and disinformation will be positively associated with apophenic constructs (bullshit receptivity, ontological confusion, conspiracy beliefs and distrust in medicine) and positively associated with cognitive reflection.

The second sub-study used a single-session, cross-sectional design and confirmatory factor analysis (CFA) to formally test the factorial structure identified in the exploratory phase. The central hypothesis was that the factor structure identified in the exploratory analysis would demonstrate good model fit.

The third sub-study was designed as a replication and extension of the confirmatory model in an independent sample. In addition to testing the stability of the factorial structure, this sub-study incorporated theoretically relevant variables (e.g., cognitive reflection, apophenic traits, and news media literacy) to assess convergent validity. Furthermore, it extended the analysis to pandemic-related behavioral outcomes, specifically the use of protective measures and vaccination intent, in order to examine the real-world relevance of fake news vulnerability. The first hypothesis was that the factorial structure would replicate. The second and third that higher vulnerability would be associated with lower news media literacy, lower engagement in protective behaviors and reduced vaccination intent. We expected correlations between vulnerability and theoretically relevant variables to mirror those found in sub-study 1.

Substudy 1

Method

Participants

Participants were recruited from two Romanian high schools (final-year students) and one university (first-year psychology students). This convenience sample aimed to maximize participation and minimize attrition, supported by educational staff who encouraged enrollment and provided weekly reminders.

Out of the 222 collected email addresses, 91 participants responded to the initial email, 76 completed the second assessment, 73 completed the third assessment and 71 participants (23 male, 48 female, m_{age} = 18.21; sd_{age} = 1.80) completed the final assessment.

Materials

As noted earlier, the headlines-only format may not provide enough detail for participants to discern intent to harm, so we developed a fake news vulnerability measure using full-text articles. Each article consisted of four paragraphs: an introduction, a section detailing the intervention (i.e., therapeutic package), a theoretical explanation, and a conclusion, designed to mirror actual health-related news. No additional information beyond the title and the specified paragraphs was included.

Sources were chosen based on Berezow's (2017) analysis, which categorized science and health news outlets by trustworthiness and story quality. For unreliable sources, we included those described by Novella (2010) as promoting "medical nonsense". Articles were selected on ten health topics: depression, allergies, dementia, HIV/AIDS, gout, pneumonia, asthma, infertility, and stress (similar to Pennycook et al., 2020). Reliable articles were sourced from trustworthy outlets, while those from unreliable outlets came from Berezow's "pure garbage" category. For three topics, we used the "official inspiration generator for alternative medicine" (Denayer, n.d.) to create multi-paragraph articles using buzzwords from

complementary and alternative medicine. These generated articles fit the health-related bullshit category (Pennycook et al., 2015).

To ensure consistent article lengths, paragraphs were minimally altered, removing or adding information as needed. Articles were categorized as misinformation or disinformation based on the use of outrage language, attacks on conventional medicine, or suggestions to discontinue treatment. Misinformation versions omitted these elements, while disinformation versions included similar elements sourced from other articles. Disinformation versions were also generated for bullshit articles. Each health topic had three versions: reliable information, misinformation, and disinformation.

Participants rated articles on four health topics, with half sourced from untrustworthy outlets and the other half randomly generated (see Table 1) and rated the trustworthiness of each article on a 5-point scale from 'Very low' to 'Very high.' A trustworthy article was defined as containing accurate, reliable information. Misinformation and disinformation scores were calculated by summing trust ratings for articles in each respective category. Fake news and health-related bullshit vulnerability scores were based on ratings of articles from disingenuous sources and randomly generated content. The information disorder score combined misinformation and disinformation scores, with higher scores indicating greater vulnerability to fake news or bullshit. The real news score was calculated by summing ratings of articles from reputable outlets, and media discernment was determined by subtracting the average information disorder score from the average real news score. A higher media discernment score reflected a stronger ability to distinguish between real and fake news.

Table 1. *News article sources and manipulations*

Topic	Reliable information	Misinformation	Disinformation
Asthma	LiveScience[1]	Randomly generated[1]	Added manipulation[1, 2]
Infertility	LiveScience	Randomly generated	Added manipulation[2]
ADHD	ScienceDaily	Randomly generated	Added manipulation

Stress	Science News[1]	Randomly generated[1, 2]	Added manipulation[1]
Gout	ScienceDaily[1]	Added manipulation[1]	Natural News[1, 2]
Pneumonia	The Guardian	Food Babe[2]	Natural News
HIV/AIDS	ScienceDaily[1, 2]	Natural News[1]	Added manipulation[1]
Depression	ScienceDaily[2]	Added manipulation	Natural News
Dementia	ScienceDaily[2]	Added manipulation	Natural News
Allergies	ScienceDaily	Natural News[2]	Added manipulation

[1] Used in Study 1
[2] Used in Studies 2 and 3
See https://osf.io/w3sq9 for full texts

Pseudo-profound bullshit receptivity (BR) was assessed using the 40-item scale designed by Pennycook et al. (2015). It features 30 bullshit sentences and 10 conventionally profound quotes. The perceived profundity of both types of items were rated on a scale ranging from '1 - not at all profound' to '5 - very profound'. A bullshit receptivity score (BRS) was calculated by summing the profundity ratings of bullshit items (Cronbach's α = .94). Bullshit sensitivity (BDS) was computed by subtracting averaged BR scores from the averaged sum of scores of the conventionally profound quotes. The Romanian translation of the BRS has shown similar characteristics and patterns of relating to other constructs as the original (Čavojová et al., 2019).

Cognitive reflection was measured using the seven-item version of the Cognitive Reflection Test (CRT - Toplak, West, and Stanovich, 2014). It retains the basic structure of the previous version (Frederick, 2005), presenting logic problems that cue an incorrect intuitive response which needs to be verified and disregarded in order to reach the correct answer. The measure presented good internal consistency (Cronbach's α = .81).

The scale designed by Lindeman (2011) was used to measure belief in CAM. The scale presents twelve methods (e.g., homeopathy, spiritual healing, oriental medicine) and asks the participants to rate how much they believe in them on a 5-point scale, ranging from '1 - do not believe at all' to '5 - fully believe', with the option '0 - cannot say'. It showed good internal consistency (Cronbach's α = .87)

Conspiracy ideation was assessed using the Generic Conspiracist Beliefs scale (Brotherton, French, and Pickering, 2013), a measure which contains 15 conspiratorial affirmations. Participants are asked to assess them on a 5-point scale ranging from ‘Definitely not true’ to ‘Definitely true’. High scores indicate pronounced conspiracy ideation. The scale presented excellent internal consistency (Cronbach's $\alpha = .91$).

Distrust in conventional medicine (DIS) was assessed using the following two questions: (1) “How much do you trust the medical system ?”; (2) “How much do you trust conventional medicine?”. The questions were rated on a 4-point scale ranging from ‘1 - Complete distrust’ to ‘4 - Complete trust’. A general score was computed by reverse-coding the two items and averaging their ratings. High scores indicate distrust in conventional medicine.

Procedure

To eliminate the possibility that differences in misinformation and disinformation vulnerability scores are based on the pairing between topic and version, we employed a repeated measure format. Over three weeks, participants rated different versions (reliable, misinformation, disinformation) of articles on the previously mentioned four health topics, completing short scales between ratings to minimize the likelihood of guessing the study's true aim. Each participant evaluated a total of 12 articles. All materials were presented in Romanian.

A cover story was used to prevent participants from detecting the true research aim. During a brief oral presentation, participants were informed that the study aimed to assess the effect of reader feedback on health-related articles written by journalism students, who would revise the articles based on this feedback. Participants were instructed to evaluate each article independently of previous versions. Email addresses were collected from interested participants, and study participation links were sent via email.

Results

An exploratory factor analysis (EFA) using principle axis factoring was applied on the 12 items measuring health-related fake news vulnerability. The suitability of the EFA was assessed prior to analysis. Inspection of the correlation matrix showed that all variables had at least one correlation coefficient greater than .3. The overall Kaiser-Meyer-Olkin (KMO) measure was .83 with individual KMO measures all greater than .61, having classifications of 'meritorious' to 'marvellous' according to Kaiser (1974). Bartlett's Test of Sphericity was statistically significant ($p < .001$), indicating that the data was likely to factorize.

The EFA revealed two factors that had eigenvalues greater than one and which explained 30.48% and 12.16% of the total variance. Visual inspection of the scree plots indicated that two factors should be retained (Cattell, 1966).

The two-factor solution explained 42.65% of the total variance of trustworthiness scores. A Promax oblique rotation was employed as factors were expected to correlate. The rotated solutions exhibited 'simple structure' (Thurstone, 1947). The interpretation of the data was partially consistent with the dimensions the scales were designed to measure, with strong loadings of real news items on Factor 2, and fake news and bullshit items, not differentiating between misinformation or disinformation, on Factor 1. Factor loadings of the rotated solution are presented in Table 2.

Table 2. *Rotated Pattern Matrix*

	Factor	
	1	2
AIDS D	**.854**	-.148
Asthma D	**.706**	.065
Gout D	**.610**	-.083
Stress D	**.300**	.164
Gout M	**.719**	-.167
Stress M	**.664**	.041
Asthma M	**.607**	.349

AIDS M	**.561**	.221
AIDS RN	-.159	**.824**
Gout RN	-.033	**.697**
Stress RN	-.007	**.458**
Asthma RN	.072	**.403**

Note. Major loadings for each item are bolded, stress and asthma M, D versions were randomly generated.

Since the EFA revealed no differences among fake news, health-related bullshit, misinformation, and disinformation, we will present only the combined data, referred to as "information disorder," in the following sections. Detailed analyses for each category are available in the Supplementary Materials.

The information disorder scale presented good internal consistency (Cronbach's $\alpha = .84$) while the real news scale presented acceptable internal consistency (Cronbach's $\alpha = .67$). No item removal would lead to improved consistency for either scale. There was a significant correlation between the two scores ($r = .25$, $p = .031$).

The articles collected from reliable news outlets received a mean trustworthiness rating of "high level of trust" ($m = 4.06$, $sd = .65$), while the other articles received a mean rating of "medium level of trust" ($m = 3.12$, $sd = .78$). The difference was statistically significant ($t = 9.03$, $p < .001$, Cohen $d = .87$).

Trust in real news was correlated positively with cognitive reflection ($r(73) = .25$, $p = .032$) and negatively with distrust in medicine ($r(73) = -.31$, $p = .009$). Information disorder vulnerability was correlated positively with pseudo-profound bullshit receptivity ($r(71) = .51$, $p < .001$), conspiracy beliefs ($r(71) = .36$, $p = .002$), belief in complementary and alternative medicine ($r(71) = .42$, $p = .009$) and distrust in medicine ($r(73) = .30$, $p = .009$). Media discernment was significantly correlated with all relevant variables (see Table 3).

Table 3. *Descriptive values and zero-order correlations*

	Mean (SD)	BRS	BDS	CRT	GCB	CAM	DIS
Real news	4.06 (.65)	.12	.20	.25*	.05	.09	-.31**
Information disorder	3.12 (.78)	.52**	-.14	-.06	.36**	.42**	.30**
Media discernment	.93 (.88)	-.37**	.27*	.24*	-.29*	-.31**	-.49**
Mean		3.42	86.01	35.9	18.56	43.55	.72
(SD)		(2.39)	(21.74)	(8.80)	(11.00)	(12.57)	(.93)
N		71	71	91	71	71	73

Note. BRS = Pseudo-profound Bullshit Receptivity, BDS = Pseudo-profound Bullshit Sensitivity, CRT = Cognitive Reflection, GCB = Conspiracy Ideation, CAM = Belief in Complementary and Alternative Medicine, DIS = Distrust in Medicine, * $p < .05$, ** $p < .01$.

Discussion

Participants were able to differentiate between trustworthy and untrustworthy articles, confirming the success of the article selection process. Although disinformation articles were rated slightly less trustworthy than misinformation (*mean difference* = .29, *SE* = .08, $p = .006$, 95%CI [.05, .53], Cohen's $d = .44$, see Supplementary Materials for full analysis), the exploratory factor analysis revealed that the categories proposed by the Information Disorder framework do not represent distinct latent factors. This implies that while the distinction between misinformation and disinformation is important for educators and policymakers, it may not be reflected in cognition and behavior.

Despite expected associations with relevant variables, repeated exposure to the articles in our design may have influenced the trust scores, in line with the illusory truth effect (Pennycook et al., 2018) which suggests that previous exposure to information results in it becoming more believable. Although we found no order effects (see Supplementary Materials), this may be due to limited statistical power.

Substudy 2

To address the limitations of Study 1, we used a single-measure format and employed Confirmatory Factor Analysis (CFA) to test and validate the proposed model while minimizing the influence of repeated exposure. Additionally, we included articles on more varied health topics (see Table 1) and extended our population to older adults.

Method

Participants

A sample of 261 Romanian participants was recruited through advertising on social media. The age range was between 18 and 68 years (m = 29.47, sd = 12.25). Female participants accounted for 77.4% of the sample, male participants for 21.5% and non-binary identified participants for 1.1%. Regarding the highest level of education achieved, 41% indicated having completed high school, 31.4% obtaining a bachelor's degree and 27.6% obtaining a master's degree or higher.

Materials

Nine articles on various health topics were selected from the database to assess fake news vulnerability in a single session. Three articles were presented in each version (reliable news, misinformation, disinformation), with an equal number of fake news and bullshit items (see Table 1 for details on topics, versions, and article types). No additional measures were included to aid in the recruitment of a large enough sample for the required statistical analyses.

Procedure

The study invitation was posted on general Facebook and Reddit groups (e.g., national, city, psychology groups). Participants were encouraged to share the study further and were offered a chance to win one of ten €20 vouchers. The invitation introduced an emerging news platform seeking to understand how readers engage with various styles and health-related topics. The nine articles were presented in random order for each participant. After concluding data collection, participants were debriefed through email about the true aims of the study.

Results

Confirmatory factor analysis, in AMOS v20 using maximum likelihood estimation, was employed to assess the two-factor model identified in Study 1 as well as two other models based on our initial theoretical approach. Acceptable model fit was evaluated based on Hu and Bentler (1999): SRMR ≤ .08, RMSEA ≤ .06, CFI and TLI ≥ .95. Factors were scaled using Unit Variance Identification (UVI) and based on the results from Study 1, they were expected to correlate. The models are presented in Figure 1.

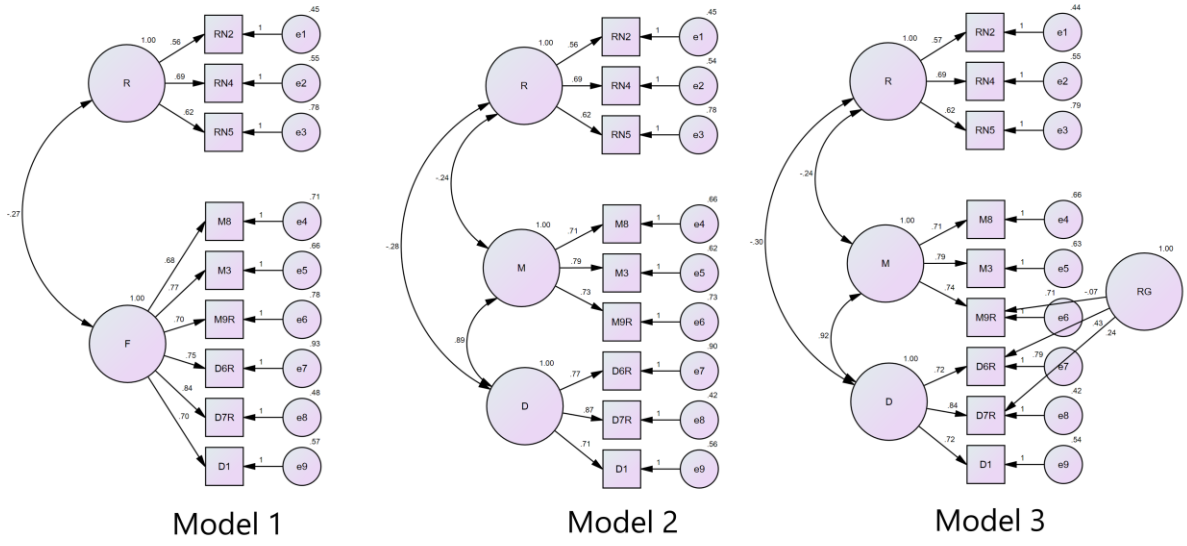


Figure 1. Tested models

The Study 1 model fitted the data in an acceptable manner ($\chi^2(26) = 55.73$, $p = .001$, TLI = .93, CFI = .95, RMSEA = .06, 90%CI [.04, .09], SRMR = .05, AIC = 93.73). Next, we tested a three-factor model which differentiates between misinformation and disinformation vulnerability, consistent with our initial theoretical approach (Wardle and Derakhshan, 2017). The model provided an acceptable fit ($\chi^2(24) = 47.37$, $p = .003$, TLI = .94, CFI = .96, RMSEA = .06, 90%CI [.04, .09], SRMR = .05, AIC = 89.37). Finally, we tested a bifactor model which also included the method of generation as a latent variable. The model fitted the data in an acceptable manner ($\chi^2(21) = 47.37$, $p = .023$, TLI = .94, CFI = .96, RMSEA = .07,

90%CI [.04, .09], SRMR = .05, AIC = 92.22). Both model 2 and 3 passed the χ^2_{diff} test (see Table 4). However, item loadings on the method factor introduced in model 3 were not statistically significant (all *p*'s > .531) and the correlation between misinformation and disinformation vulnerability in model 2 was .89, which exceeds the .85 cut-off score for problematic discriminant validity (Cohen et al., 2003). Hence, the initial model was considered to best fit the data. One area of ill fit was identified by analysing the standardized residual covariance matrix, related to the reliable article on the topic of depression and the disinformation article about gout (*z* = -3.30). All parameter estimates were statistically significant (all *p*'s < .001, See Supplementary Materials), there were no out-of-range values, negative factor variances or negative indicator error variances.

Table 4. *Fit indices for the alternative models*

Item	Model 1	Model 2	Model 3
χ^2	55.726	47.37	44.22
df	26	24	21
p	.001	.003	.002
TLI	.933	.943	.935
CFI	.952	.962	.962
RMSEA	.066	.061	.065
90% CI	.042, .090	.035, .087	.038, .092
SRMR	.0525	.0502	.0489
AIC	93.726	89.37	92.22
χ^2_{diff}		8.356	3.15

The subscales presented acceptable internal consistency: Cronbach's α = .72 for the Misinformation subscale, Cronbach's α = .74 for the Disinformation subscale, Cronbach's α = .70 for Health-related bullshit subscale, Cronbach's α = .70 for the Fake news subscale, Cronbach's α = .66 for the Real news subscale, Cronbach's α = .83 for the Information disorder subscale.

The articles collected from reliable news outlets received a mean trustworthiness rating of "high level of trust" (*m* = 3.78, *sd* = .77), while the other articles received a mean rating of

"low level of trust" (*m* = 2.30, *sd* = .85 for the those collected from unreliable sources; *m* = 2.61, *sd* = .91 for randomly generated articles; *m* = 2.64, *sd* = .89 for misinformation; *m* = 2.27, *sd* = .91 for disinformation; *m* = 2.46, *sd* = .81 for information disorder).

Paired samples t tests were employed to investigate the difference between reliable news and information disorder vulnerability, misinformation and disinformation, and fake news and health-related bullshit. The difference between reliable news and the other collected articles was statistically significant and indicated a very large effect size ($t(260) = 17.46$, $d = 1.11$, 95%CI [.93, 1.30]). Medium to large, statistically significant difference were observed between misinformation and disinformation vulnerability ($t(260) = 7.81$, $d = .49$, 95%CI [.32, .66]) and between fake news and health-related bullshit vulnerability ($t(260) = 7.54$, $d = .48$, 95%CI [.31, .66]).

Discussion

Study 2 confirmed the findings of the first study. The two-factor model provided the best fit, with high correlation between misinformation and disinformation vulnerabilities, suggesting these categories are not influenced by distinct cognitive processes. The lack of significant findings in the bifactor model suggests that how information is generated (unreliable sources vs. random generation) does not substantially affect trustworthiness ratings.

Substudy 3

Given the consistency of results between the single-session and repeated measures in the previous studies regarding factorial structure, internal consistency, and trust in the articles, we designed a two-wave study to explore the relationship between fake news vulnerability and other relevant constructs. When the COVID-19 pandemic began shortly after we started recruitment, we added a third set of measures to assess the behavioral implications of fake news vulnerability, focusing on pandemic-related variables. This battery also measured news

media literacy, a key target of fake news interventions (Ireton and Posetti, 2018; Basol et al., 2020). In March 2021, with the start of Romania's vaccination rollout, we sent a final question regarding participants' vaccination intent.

Method

Participants

A total of 180 Romanian participants completed the first testing phase, 139 completed the second testing phase and 97 completed the third testing phase (ages 18-78, m = 29.77, sd = 13.44, 24 males, 73 females). The final testing phase was completed by 65 participants (ages 18-78, m = 29.57, sd = 13.40, 13 males, 52 females).

Participants were recruited through Facebook groups and Reddit posts, with snowball sampling used to reach a diverse audience. Twenty-six participants had medical school experience. Of the sample, 5.6% had not completed high school, 43.9% had a high school diploma, 33.9% held an undergraduate degree, and 16.7% had a graduate degree.

Materials

We employed the single-session fake news vulnerability assessment from Study 2, along with the measures from Study 1. The only change was the removal of the middle point (i.e., '3') from the BRS, GCB, and CAM scales to prevent it being used as a "don't know" option (Mækelæ et al., 2018). All scales were rated on 4-point Likert scales.

News Media Literacy was assessed using The News Media Literacy Scale (NML; Ashley et al., 2013). The scale assesses knowledge related to news media industry, content and effects by presenting participants with 14 statements which they were asked to rate on a scale between '1 - strongly disagree' and '7 - strongly agree'.

Participants reported how many hours they spent consuming news and which sources they used in the past week. Options included television, radio, written press, websites, social media, online platforms, and relatives or friends. Responses were combined to quantify the variety of news sources used.

Attitudes towards the intentional, anthropogenic origin of the SARS-COV-2 virus were assessed using a visual-analog scale (VAS) marked with “certainly a product of natural phenomena” and “certainly a product of human engineering”. Responses were converted to a 0-100 metric, lower scores indicating preference for natural origins.

The behavioral influence of fake news was measured by asking participants which protective measures they had used at least once against COVID-19 (see Table 5). "Actual protective measures" were those recommended by authorities (e.g., hand-washing, mask-wearing, social distancing), while "fake protective measures" were promoted through misinformation (e.g., drinking water frequently, prayer, immunity-boosting supplements). Participants could also submit other measures, which were categorized as actual or fake based on WHO guidelines. Three indices were analyzed: the number of actual protective measures, fake protective measures, and the Protective Measure Discrimination Index (PMDI), calculated by subtracting fake from actual measures.

Table 5. Use of preventive measures

Actual protective measures	**Number of participants (%)**
frequent hand-washing	94 (96.91)
disinfecting hands	83 (85.57)
reducing social contacts	64 (65.98)
eliminating social contacts	27 (27.84)
avoiding face touching in public places	75 (77.32)
wearing a mask	59 (60.82)
avoiding public transportation	56 (57.73)
disinfecting your home	44 (45.36)
keep at least 1m apart from other people	70 (72.16)

Fake protective measures	
frequent water drinking (e.g., once every 15 minutes)	33 (34.02)
frequent hot beverage drinking	13 (13.40)
religious protective measures (e.g., prayer, communion)	15 (15.46)
immunity boosting through supplements	41(42.27)
preventively drinking alcohol	5 (5.15)
spiritual protective measures (e.g., energetic cleansing)	14 (14.43)
inhaling steam to cleanse respiratory airways	3 (3.09)
Other	**Actual/Fake**
disinfecting groceries (1 participant, 1.03%)	Fake
wearing single use gloves (2 participants, 2.06%)	Actual
eating more fruits and vegetables to boost immunity (1 participant, 1.03%)	Fake
stimulating immunity through sport, yoga (1 participant, 1.03%)	Fake
disinfecting desk and laptop at work (1 participant, 1.03%)	Actual

Vaccination intent was assessed with the question: "Do you plan to get vaccinated against COVID-19?" on a 5-point scale from 1 (Definitely not) to 5 (Definitely yes).

Procedure

Study recruitment began in November 2019, with invitations posted on Facebook and Reddit. The objective was framed as examining the relationship between psychological variables and responses to health-related news and profound quotes, encouraging participants to share the invitation. Due to the length of the assessment, it was conducted in two phases. In the first phase, participants rated 9 news articles and completed the BRS, with introduced with the following coverup story: "We would like to assess the quality of a few journalistic articles on health-related topics, so that we can offer feedback to the authors based on the psychological characteristics of their raters". The nine items were presented in a random order for each participant.

A week after completing the first phase, participants were invited via email to complete the second phase, which included the CRT, CAM, GCB, and DIS measures. Clinical measures were included and will be presented elsewhere. Upon completing both phases, participants

could enter a lottery for one of ten €20 vouchers. Weekly social media reposts continued until February 2020.

In March 2020, after the national lockdown, participants who completed the second phase were invited to a third phase, which included the News Media Literacy Scale, news platform usage, time spent on news, healthcare professional status, virus origin (VAS). Participants were also asked to select the protective behaviors they employed against COVID. Another lottery for €20 vouchers was offered. In March 2021, a final question on vaccination intent was sent via email to participants who completed the third phase to encourage maximum response amid high attrition.

Results

Confirmatory factor analysis, in AMOS v20 using maximum likelihood estimation, was employed to assess the two-factor model identified in Study 1 and 2. The model fitted the data in an acceptable manner ($\chi^2(26) = 36.37$, $p = .085$, TLI = .97, CFI = .98, RMSEA = .05, 90%CI [.00, .08], SRMR = .05). No areas of ill fit were identified by analysing the standardized residual covariance matrix (all z's < 1.79). All parameter estimates were statistically significant (all p's < .001, See Supplementary Materials), there were no out-of-range values, negative factor variances or negative indicator error variances.

The subscales presented acceptable internal consistency: Cronbach's $\alpha = .60$ for the Real news subscale, Cronbach's $\alpha = .86$ for the Information disorder subscale.

The articles collected from reliable news outlets received a mean trustworthiness rating of "medium level of trust" ($m = 3.86$, $sd = .72$), while the other articles received a mean rating of "low level of trust" ($m = 2.06$, $sd = .82$). The difference between real news scores and information disorder scores was statistically significant ($t(179) = 21.59$, $p < .001$, Cohen's d

= 1.72, 95%CI[1.48; 1.96]). Detailed trustworthiness ratings and zero-order correlations can be seen in Table 6.

Table 6. *Descriptive values and Pearson correlations*

	Mean (SD)	BRS	BDS	CRT	GCB	CAM	DIS	NML	VNPS	DAT	CoVO	APM	FPM	PMDI	VI
RN	3.86 (.72)	-.05	.10	.11	-.17*	-.10	-.22**	.23*	.16	-.14	-.18	-.06	-.02	-.05	.44**
ID	2.06 (.82)	.48**	-.11	-.36**	.53**	.52**	.38**	-.11	-.26*	-.07	.25*	-.15	.15	-.20*	-.32*
MD	.90 (.56)	-.39**	.15*	.35**	-.52**	-.47**	-.44**	.23*	.30**	-.03	-.30**	.08	-.13	.13	.53**
	Mean	29.68	.56	3.89	28.91	14.48	2.94	88.00	2.67	2.00	33.11	5.48	.57	4.92	2.74
	(SD)	(9.96)	(.59)	(2.21)	(10.42)	(7.64)	(.71)	(15.3)	(1.07)	(1.76)	(31.32)	(1.7)	(.68)	(1.77)	(1.50)
	N	180	180	139	139	139	139	97	97	97	97	97	97	97	65

Note. RN = Reliable News, FN = Fake News, HBS = Health-related Bullshit, M = Misinformation, D = Disinformation, ID = Information Disorder, MD = Media Discernment, BRS = Pseudo-profound Bullshit Receptivity, BDS =Pseudo-profound Bullshit Sensitivity, CRT = Cognitive Reflection, GCB = Conspiracy Ideation, CAM = Belief in Complementary and Alternative Medicine, DIS = Distrust in Medicine, NML = News Media Literacy, VNPS = Variety of News Platform Sources, DAT = Daily Average Time reading news, CoVO = Virus origin, APM = Actual Protective Measures, FPM = Fake Protective Measures, PMDI = Protective Measures Discrimination Index, VI =Vaccination Intent, * $p < .05$, ** $p < .01$.

Trust in real news was negatively associated with conspiracy ideation ($r(139) = -.17$, $p = .040$), distrust in conventional medicine ($r(139) = -.22$, $p = .008$), and positively associated with news media literacy ($r(97) = .23$, $p = .024$) and vaccine intention ($r(65) = .45$, $p < .001$).

Information disorder vulnerability was positively associated with pseudo-profound bullshit receptivity ($r(180) = .40$, $p < .001$), conspiracy ideation ($r(139) = .53$, $p < .001$), belief in CAM ($r(139) = .53$, $p < .001$), distrust in conventional medicine ($r(139) = .38$, $p < .001$), and considering COVID to be artificially produced ($r(97) = .25$, $p = .016$) and negatively associated with cognitive reflection ($r(139) = -.36$, $p < .001$), number of news sources used ($r(139) = -.26$, $p = .009$), the Protective Measure Discrimination Index (PMDI, $r(97) = -.20$, $p = .045$) and vaccination intent ($r(65) = -.31$, $p = .011$).

Media discernment was negatively associated with pseudo-profound bullshit receptivity ($r(180) = -.39$, $p < .001$), conspiracy ideation ($r(139) = -.52$, $p < .001$), belief in CAM ($r(139) = -.47$, $p < .001$), distrust in conventional medicine ($r(139) = -.44$, $p < .001$), and considering COVID to be artificially produced ($r(97) = -.30$, $p = .003$), being positively associated with pseudo-profound bullshit sensitivity ($r(180) = .15$, $p = .046$), cognitive reflection ($r(139) =$

.35, $p < .001$), number of news sources used ($r(97) = .30$, $p = .002$), news media literacy ($r(97) = .23$, $p = .026$), and vaccination intent ($r(65) = .52$, $p < .001$).

The participants that reported being healthcare workers or studying to become healthcare workers offered significantly lower trustworthiness scores than the rest of the sample to information disorder articles ($t(56.18) = 3.50$, $p = .001$, $d = .50$, 95%CI [.09, .92]), also having higher media discernment scores ($t(41.52) = 2.83$, $p = .007$, $d = .59$, 95%CI [.07, .91]). However, there were no significant differences between them in terms of reliable news trustworthiness scores ($t(178) = .84$, $p = .401$, $d = .18$, 95%CI [-.24, .59]).

Discussion and conclusions

This research explored distinctions between misinformation, disinformation, and health-related bullshit within the Information Disorder framework, focusing on cognitive and behavioral impacts during the COVID-19 pandemic. Across three studies, we examined participants' vulnerability to various forms of health-related fake news and bullshit, their ability to discern trustworthy from untrustworthy information, and the role of psychological factors such as cognitive reflection, conspiracy ideation, and belief in complementary and alternative medicine (CAM) in shaping this vulnerability. We also investigated how these factors influenced protective behaviors and vaccination intent during the pandemic.

The two-factor structure identified in Study 1 was confirmed in Study 2 and 3. Both exploratory and confirmatory factor analyses differentiated between reliable and unreliable articles, without differentiating between disinformation and misinformation or between health-related fake news and bullshit (i.e., randomly generated content). The distinction between misinformation and disinformation is important for policy and educational interventions on a theoretical and didactical level as argued by Wardle and Derakhshan (2017). However, our results suggest that this distinction does not reflect the cognitive

processes that make people vulnerable to fake news, that there is no separate vulnerability towards misinformation and disinformation. Tandoc et al. (2018) showed that the term "disinformation" evokes less concern, perceptions of falsity and intentionality than the term "misinformation", which is at odds with the information disorder framework. Our results indicate that while the term might not evoke more concern, content specific to health-related disinformation (i.e., outrage language, attacks on conventional medicine, or suggestions to discontinue treatment) is invested with less trust. This medium-to-large significant difference was observed in all three samples, across different ages and article topics. The utility of the information disorder framework is evident in training people to reliably identify these elements, observe their impact and classify the articles as disinformation. From a conflict detection perspective (De Neys, 2014), having people engage in effortful, controlled, critical thinking will be useful regardless of the type of content they encounter.

An important clarification of these findings is that the emergence of a common factor should not be interpreted as evidence that individuals ignore or are incapable of evaluating producer intent. Rather, the results indicate that the same individuals who are vulnerable to misinformation are also vulnerable to disinformation, suggesting a shared underlying susceptibility that cuts across content types. In practical terms, this means that vulnerability is not selective: individuals who tend to accept inaccurate information are also more likely to accept intentionally deceptive information when it is encountered. From a cognitive standpoint, this points to a general disposition toward trusting misleading content under conditions of fluency, coherence, or affective appeal, rather than to separate evaluative mechanisms tuned to different kinds of falsehood.

While the distinction between the two informational disorder categories remains essential for regulatory, educational, and ethical purposes, the present results suggest that interventions targeting vulnerability must address the common cognitive and behavioral mechanisms that

underlie both. Training individuals to recognize intent alone is unlikely to be sufficient if the same processing biases and trust allocation strategies apply across content types. Instead, reducing fake news vulnerability requires strengthening conflict detection, improving sensitivity to concrete manipulation cues, and disrupting feedback loops that expose the same individuals to progressively more misleading information. In this sense, the identification of a shared factor highlights a need for interventions that operate at the level of underlying susceptibility rather than content classification alone. This does not signal a failure of conceptual distinctions.

News media literacy, defined as the ability to access, evaluate, and analyze news media content (Ashley et al., 2013), has been a key focus in efforts to combat the spread of fake news, with several interventions targeting it (e.g., Ireton and Posetti, 2018; Basol et al., 2020). However, Jones-Jang et al. (2021) found that neither news, media, nor digital literacy predicted media discernment, instead identifying information literacy as the most effective predictor. Our findings indicate that news media literacy is more strongly associated with trust in reliable news than with trust in unreliable news ($z = 1.74$, $p = .041$), highlighting its selective influence in promoting trust in credible sources.

A key aspect of news media literacy is encouraging users to cross-check sources and diversify the platforms they use. In Study 3, the variety of news sources was correlated with both lower information disorder vulnerability and higher media discernment. Interestingly, the total daily time spent consuming news was not linked to information disorder vulnerability. This suggests that how individuals engage with news, in terms of source diversity and evaluative practices, may be more consequential for fake news vulnerability than sheer exposure volume.

Both the use of actual COVID-19 protective measures and fake protective measures showed the expected associations with information disorder vulnerability and media discernment, albeit these associations did not reach statistical significance. The protective measure discrimination index, which was computed by subtracting the number of fake measures from the number of actual measures used, revealed a small but significant negative correlation with fake news vulnerability. While recent studies have found mixed results regarding the links between conspiracy beliefs related to COVID-19, knowledge about protective measures, and their use (Lee et al., 2020; Allington et al., 2021; Naveed et al., 2021), our data suggests a more nuanced relationship. Specifically, belief in the anthropogenic origin of COVID-19 was associated with the use of fake prevention measures ($r(97) = .23$, $p = .024$) but not with the use of actual prevention measures ($r(97) = .07$, $p = .490$). Furthermore, general conspiracy ideation did not correlate significantly with the use of either protective measure category ($r(97) = .02$, $p = .812$ and $r(97) = .06$, $p = .586$). Although general conspiracy ideation showed a small correlation with belief in the anthropogenic origin of COVID-19 ($r(97) = .25$, $p = .016$), our findings suggest that COVID-19-specific conspiracy beliefs are more strongly linked to information disorder vulnerability than to pre-existing conspiracy ideation. This underscores the critical role that misinformation plays in influencing public health behaviors during a crisis.

Loomba et al. (2021) experimentally demonstrated the impact of misinformation on vaccination intent. Our data provide longitudinal evidence supporting this effect, as participants evaluated news articles roughly one year prior to the assessment of their vaccination intent. Vaccination intent showed medium, positive associations with trust in reliable news and media discernment, while demonstrating small to medium negative correlations with fake news vulnerability. This suggests that greater trust in credible sources and stronger media discernment skills are linked to a higher likelihood of intending to get

vaccinated, whereas vulnerability to fake news reduces this intent, further underscoring the detrimental role of misinformation in shaping health-related behaviors.

Taken together, these findings suggest that the relationship between information disorder vulnerability and health behavior appears to manifest primarily in the selection of prevention strategies rather than reflecting a generalized tendency to reject protective behaviors. Individuals who are more vulnerable to information disorder are not necessarily less active in attempting to protect themselves, they are more likely to adopt ineffective or unfounded measures. This distinction is important from a behavioral standpoint, as it indicates that misinformation does not promote passivity or disengagement. It can however actively redirect behavior toward maladaptive actions that provide an illusion of control or agency during periods of uncertainty.

The dissociation between general conspiracy ideation and actual health behavior further supports this interpretation. Broad conspiratorial thinking does not appear to directly translate into concrete behavioral choices unless it is instantiated in context-specific narratives. COVID-19–related conspiracy beliefs, by contrast, were meaningfully associated with the use of fake protective measures, suggesting that situationally relevant misinformation exerts a stronger influence on behavior than abstract conspiratorial worldviews. In practical terms, this implies that during public health crises, exposure to domain-specific misinformation may override or bypass more general belief systems, shaping behavior through immediate relevance and perceived applicability rather than through longstanding ideological commitments.

Finally, the longitudinal association between fake news vulnerability and vaccination intent highlights the lasting behavioral consequences of earlier exposure to misinformation. Even when beliefs are assessed at some time after initial exposure, vulnerability to misleading

content remains linked to reduced trust in reliable sources and lower willingness to engage in protective health behaviors. This persistence underscores the importance of early intervention, as misinformation encountered during critical periods may shape downstream decisions well beyond the immediate context. From a public health perspective, these findings suggest that combating misinformation is not only about correcting false beliefs, but about preventing the gradual erosion of trust and discernment that ultimately influences real-world health decisions.

Limitations and further research

In all three studies, convenience sampling was employed, meaning the data cannot be considered representative of the Romanian population, and the samples were not gender balanced. This asymmetry warrants consideration when interpreting the findings, as prior research has documented gender differences in several domains relevant to fake news vulnerability. For instance, women tend to report higher engagement with health-related information and greater concern for health risks, particularly in the context of public health crises (Ek, 2015; Galasso et al., 2020), which may influence baseline responses to health-related news content. The overrepresentation of women may therefore have contributed to stronger engagement with pandemic-related materials or heightened sensitivity to health messaging. Greater exposure to health information and higher perceived relevance of the topic may have facilitated more informed judgments in some cases, potentially increasing sensitivity to misleading claims. At the same time, higher involvement and prior knowledge can also introduce systematic biases, such as stronger reliance on pre-existing beliefs or source-based heuristics, which may shape trust evaluations independently of content accuracy. As a result, increased informedness may operate as both a protective factor and a biasing influence, enhancing discrimination in some contexts while reinforcing selective trust in others. Another potential limitation that can be derived from this difference is related to

differences in how trustworthiness was evaluated, limiting generalizability of the findings about the set of articles used. Importantly, however, the primary analyses focused on structural relationships among fake news vulnerability and cognitive, personality, and literacy related variables rather than on mean-level group differences. Structural associations have been shown to be more robust to sampling imbalances than comparisons of absolute levels (Little, 2013). Moreover, the consistency of results across the three samples and their alignment with existing literature lend confidence to the study's conclusions. Nonetheless, future research should employ gender-balanced or stratified samples and test measurement invariance to determine whether the identified vulnerability profiles and latent structures generalize across gender groups.

Our investigation excluded malinformation, the third category from the Information Disorder framework, focusing solely on fake news vulnerability. Intent to harm was operationalized through the use of outrage language and attacks on conventional medicine, the most common elements identified in the articles we examined. Different manifestations of intent to harm may exist in other contexts, and future research should explore whether similar patterns emerge with different types of content.

We opted for a long article format to provide enough information for the misinformation/disinformation categorization. While participants were asked to read as they would online to preserve ecological validity, this might have compromised internal validity since it's unclear if participants fully read the articles (Pennycook et al., 2021). Further research on full-text articles could benefit from the granularity and increased precision offered by investigating reading patterns using eye-tracking devices.

While Study 3 assessed overall news consumption time, it did not differentiate between specific platforms or media types. Prior work indicates that platform-specific dynamics,

particularly social media use, may play a distinct role in shaping health-related beliefs and behaviors (Allington et al., 2021). The absence of platform-level measures limits conclusions about whether source diversity reflects active epistemic engagement or simply broader media exposure. Future research should disentangle general news consumption from platform-specific use to clarify how different media ecologies contribute to fake news vulnerability and health-related decision-making.

The use of protective measures was recorded dichotomously, with participants reporting whether they had used each measure at least once. This method does not capture the intensity of use and may have led to a ceiling effect for some measures, such as frequent handwashing, which almost all participants reported. More granular measures of behavior frequency could increase precision.

When assessing vaccination intent, we did not control for medical conditions, which could better explain reluctance than fake news vulnerability. Some participants may have expressed lower vaccination intent due to legitimate medical concerns, though the relationships we observed suggest this was not a major issue.

Lastly, the successive dropout in Study 3 raises concerns about potential biases. The remaining participants may have been more engaged and up to date with COVID-19 and vaccination news, potentially skewing results. As such, these findings may not fully generalize to the wider population, highlighting the need for further research to account for these variables, such as personality traits like conscientiousness.

Study 2. The higher you think of yourself, the harder you fall: Overconfidence as a distinct, mutable predictor of fake news vulnerability

The present study aimed to determine whether memory overconfidence serves as a predictor of health-related fake news vulnerability and to evaluate the impact of overconfidence-

correcting feedback on trust in fake news. We hypothesized that overconfidence in verbal working memory (H1) and general knowledge (i.e., long-term memory, H2) are predictive of fake news vulnerability, controlling for other reflexive and reflective open-minded thinking. As Fleming and Lau (2014) distinguish between metacognitive sensitivity (the ability to distinguish between correct and incorrect responses, H1.1 and H2.1) and metacognitive bias (the overall level of confidence expressed in incorrect trials, H1.2 and H2.2), we tested the predictive power of both measures.

The inattention-based account of misinformation sharing (Pennycook et al., 2021b) stipulates that people distribute fake news on social media because their behavior is controlled by alternative reinforcers (e.g., social validation) rather than not caring about the accuracy of what they share. Presenting a nudge towards accuracy, such as asking people to rate the accuracy of a headline, has been shown to reduce fake news sharing in lab experiments and on Twitter (Pennycook et al., 2021b). Providing overconfidence-correcting feedback (e.g., informing people of their overconfidence scores) can be considered a type of accuracy nudge and has been proposed as a potential intervention to reduce trust in fake news (Lyons et al., 2021, Mirhoseini et al., 2023). As reductions in memory overconfidence could increase local and global conflict detection, we hypothesized that receiving overconfidence-correcting feedback will diminish fake news vulnerability (H3.1), overconfidence acting as a change mechanism (H3.2).

Method

The preregistration for the study can be accessed at https://osf.io/x2nzy. Open Science Framework data, materials and code can be accessed at: https://osf.io/n8fyp/

Participants

An a priori power analysis (G*Power 3.1, Faul et al., 2007) suggested that a minimum of 395 participants would be required to detect a small effect size ($f^2 = .02$) for the increase in R^2 stipulated by H1 and H2, and a medium effect size ($f^2 = .25$) for H3 ($\alpha = .05$ and power = .80). A total of 740 participants were recruited from the Babeș-Bolyai student population, which was offered extra course credits, and from the general Romanian population, using general social media posts and targeted posts in various social media groups. The study was completed by 395 participants (335 female, 56 male, 4 indicating other gender identities, $M_{age} = 20.69$, $SD_{age} = 4.29$), and according to the preregistered plan, only completers were included in the analyses. Of the 354 incomplete accounts, 196 (55.37%) dropped out after answering the demographic questions and 119 (33.62%) dropped out before completing the first post-test task. All participants who completed the study were offered a chance to win one of the ten vouchers (worth 20 Euros each). The inclusion criterion was being at least 18 years old. Recruitment efforts included weekly posting and reposting for the duration of the recruitment period. Participants who opted for vouchers provided their name and their email address and were informed that multiple entries would result in exclusion from the lottery.

In terms of educational level, 1% of the sample reported having completed middle school, 78.5% had completed high school, 15.7% had obtained a bachelor's degree and 4.8% had obtained a master's degree.

Measures

Fake news vulnerability was assessed using twelve health-related news articles, eight collected from untrustworthy news outlets (Berezow, 2017) and four randomly generated. Previous research on these articles illustrated the lack of differences between the two categories of items in terms of scores, factorial structure, and association with relevant variables, suggesting that they are representative of health-related fake news (Secară, 2018,

2019). Participants were asked to rate the trustworthiness of each item on a 5-point scale ranging from 1 - Very low level of trust to 5 - Very high level of trust and were presented with the following statement "A trustworthy article was defined as contain accurate and honest information that you consider you can rely on". One set of six items was presented at pre-test (T0) and another at post-test (T1), with the order of items presented at each time point being randomised. The sets were selected based on data from previous studies. We selected articles that received similar scores and displayed good internal consistency in groups of 6 (Cronbach's α = .86 and .80, see Secară, 2019). Acceptable internal consistency was observed for the data in this study (Cronbach's α = .76 and .74). A fake news vulnerability score at each moment was computed by averaging the ratings of the articles. Six health-related articles collected from trustworthy outlets were included, three at each time point, to mask the aim of the study and avoid artificial scepticism. Each article featured a title (e.g., How to cure tuberculosis naturally with vitamin C) and consisted of four paragraphs (about 400 words total): an introduction defining key terms (e.g., tuberculosis and its treatment), a paragraph on the underlying theory (e.g., references to books and studies, Vitamin C is the fuel for the body's own immune system), another detailing the specifics of the intervention (e.g., how much and how often), and a conclusion.

Pseudo-profound bullshit receptivity was measured using 10 items from the Bullshit Receptivity Scale (BRS, Pennycook et al., 2015). Participants were presented with 10 profound sounding randomly generated sentences (e.g., "Wholeness quiets infinite phenomena") and asked to rate their perceived profundity on a 4-point scale ranging from "not at all profound" to "very profound". Pseudo-profound bullshit receptivity was computed by averaging the profoundness scores accorded to the items of the BRS. Internal consistency was acceptable (Cronbach's α = .75).

Overclaiming was measured using the "historical names and events" and the "topics in physical sciences" subscales of the Over-Claiming Questionnaire (OCQ, Paulhus et al., 2003). Fifteen items from each category were presented (e.g., "centripetal force"), three of them being fictional (e.g., "ultra-lipid"). Participants were asked to rate their familiarity with the items on a scale ranging from 1 - "slightly familiar" to 6 - "very familiar" with the option 0 - "never heard of it". An overclaiming bias score was computed by summing the familiarity scores of the fabricated items. Internal consistency was good (Cronbach's $\alpha = .86$).

Analytic thinking was measured using the Cognitive Reflection Test (CRT, Toplak, West, and Stanovich, 2014). Participants were presented with seven logical world problems that cue an incorrect intuitive response which needs to be examined and disregarded in order to reach the correct answer. The total CRT score was obtained by counting the number of correctly solved problems.

Actively open-minded thinking was measured using the 17-item version of the AOT scale developed by Svedholm-Häkkinen and Lindeman (2017). The scale assesses the disposition to think reflectively by asking participants to rate statements such as "People should always take into consideration evidence that goes against their beliefs" or "I consider myself broad-minded and tolerant of other people's lifestyles" on a scale from ranging from 1 - "strongly disagree" to 6 - "strongly agree". A general actively open-minded thinking score was obtained by summing the scores of the items. Internal consistency was acceptable (Cronbach's $\alpha = .74$).

Overconfidence related to verbal working memory was assessed using the operation span task (Unsworth et al., 2005). Each trial included a math equation which had to be solved as fast and as accurate as possible, followed by a word which needed to be remembered. The task includes three types of practice trials: equations only, words only and mixed. After three to

seven experimental trials, participants were asked to recall the words in the correct order and assess how confident they are that their response is correct (on a scale from 0 to 100).

Overconfidence related to general information was assessed using the information subscale of the Multidimensional Aptitude Battery-II (MAB-II, Jackson, 1998; Iliescu, Glință and Ispas, 2009). Participants were presented with 26 multiple choice questions related to general knowledge. After each question, they were asked to evaluate how confident they are that their response is correct (on a scale from 0 to 10).

Two metacognitive sensitivity scores were computed for both the verbal working memory and the general knowledge tasks. The first one was computed by subtracting the sum of the confidence ratings of the correct answers from sum of the confidence ratings of the incorrect answers in the first half of each test (pretest, T0) and the second, using the same formula, from the second half of each test (post-test, T1). Two metacognitive bias scores were computed for both the verbal working memory and the general knowledge tasks. The first was computed by averaging confidence ratings across the first half of the trials (pretest, T0) and the second by averaging confidence ratings across the second half of the trials (post-test, T1).

Procedure

Participants entered the study via a link to the Gorilla online experimental platform (Anwyl-Irvine et al., 2020). There they were presented with information about the study and had to give their consent before proceeding. As revealing the aim of the study might alter their responses (e.g. increase scepticism towards the news articles), they were told that the main aim of the study was to analyse how individual differences influence the response to certain types of feedback in memory tasks. After completing demographic information, participants completed the CRT, OCQ, BRS and AOT. They then rated half of the news items and

completed half of the general knowledge and working memory tasks. After each trial, confidence was assessed using a slider. At this point, the platform randomly assigned participants to either the control or experimental group. Both groups received feedback, the control group on the time taken to complete the two tasks and the intervention group on the number of incorrect answers for each task and their average confidence in the incorrect answers. To ensure that the information was retained, participants had to type in the numbers presented in order to continue. They then completed the remainder of the general knowledge and working memory tasks, each trial followed by the confidence assessment, and then rated the remaining news articles. Once all the data had been collected, participants received an email informing them of the fake articles and their average trust scores for the reliable and unreliable articles.

Data Analysis

Hierarchical regressions were used to evaluate the predictive power of overconfidence, the criterion being fake news vulnerability at pre-test (T0). The base model included pseudo-profound bullshit receptivity, overclaiming, actively open-minded thinking and analytical thinking as predictors. In the second model, the hypothesis specific predictors (e.g., sensitivity or bias for working memory or general knowledge overconfidence at T0) were added. Each variable included as a predictor was mean-centered to avoid multicollinearity.

A one-way ANCOVA was used to test H3.1, having group as a between-subjects factor, trust in health-related fake news at T0 as a covariate and trust in health-related fake news at T1 as outcome.

All data were analysed using IBM SPSS 25.

Results

The exploratory analysis of the correlations between health-related fake news vulnerability and the previously established predictors of fake news vulnerability found no significant associations when controlling for multiple comparisons (Table 1).

Table 1. Descriptive statistics and zero-order correlations for reflexive and reflective open-minded thinking, fake news vulnerability, and overconfidence at pretest

	M	SD	1	2	3	4	5	6	7	8	9
1. CRT	3.30	2.44	_								
2. OCQ	5.98	6.02	-.02	_							
3. BSR	2.40	0.56	.03	.14	_						
4. AOT	72.78	9.26	-.01	-.07	.14	_					
5. MSM1	-74.37	22.89	.05	-.04	.05	-.05	_				
6. MSO1	-345.04	336.09	-.04	.02	.06	.00	.09	_			
7. MBM1	4.45	2.34	.06	.07	.09	-.11	.20	.03	_		
8. MBO1	58.07	25.22	.02	.14	.15	-.09	-.03	.15	.25	_	
9. FN1	2.54	0.79	.05	.03	.11	-.07	.09	.06	.16	.13	_

CRT = Cognitive Reflection Test, OCQ = Overclaiming Questionnaire, BSR = Bullshit Receptivity, AOT = Actively Open-minded Thinking, MSM1 = Metacognitive Sensitivity General Knowledge at pretest, MSO1 = Metacognitive Sensitivity Working Memory at pretest, MBM1 = Metacognitive Bias General Knowledge at pretest, MBO1 = Metacognitive Bias Working Memory at pretest, FN1 = Fake News vulnerability at pretest

Table 2. Means and standard deviations used for group comparisons

	Group					
	Experimental N = 188		Control N = 207			
	M	*SD*	*M*	*SD*	*t*	*p*
FN1	2.49	0.78	2.59	0.81	1.27	0.204
MSM1	-75.43	22.52	-73.41	23.24	0.88	0.382
MSO1	-366.53	331.94	-325.72	339.41	1.20	0.230
MBM1	4.69	2.45	4.24	2.21	-1.94	0.053
MBO1	58.24	26.53	57.91	24.03	-0.13	0.898
FN2	2.40	0.74	2.50	0.77	1.36	0.175
MSM2	-41.36	32.62	-42.02	31.75	-0.20	0.838
MSO2	-648.23	256.03	-577.84	329.32	2.38	0.018
MBM2	5.63	2.19	5.68	2.08	0.24	0.811
MBO2	39.80	32.31	47.11	32.14	2.25	0.025

FN1 = Fake News vulnerability at pretest. MSM1 = Metacognitive Sensitivity General Knowledge at pretest, MSO1 = Metacognitive Sensitivity Working Memory at pretest, MBM1 = Metacognitive Bias General Knowledge at pretest, MBO1 = Metacognitive Bias Working Memory at pretest, FN2 = Fake News vulnerability at post-test. MSM2 = Metacognitive

Sensitivity General Knowledge at post-test, MSO2 = Metacognitive Sensitivity Working Memory at post-test, MBM2 = Metacognitive Bias General Knowledge at post-test, MBO2 = Metacognitive Bias Working Memory at post-test.

As a first step in assessing the added variance explained by overconfidence, we analysed the predictive power of reflective and reflexive open-minded thinking (Model 0, see Table 3). The overall model was not statistically significant ($F(4, 390) = 1.68$, $R^2 = .02$, $p = .153$). None of the variables predicted trust in health-related fake news.

Table 3. Model 0: Multiple regression analysis for variables predicting health-related fake news vulnerability

Variable	*B*	SE_B	β	*p*
Constant	15.26	.24		< .001
Cognitive reflection	.10	.10	.05	.322
Overclaiming	.01	.04	.01	.841
Bullshit receptivity	.08	.04	.09	.069
Actively open-minded thinking	-.03	.03	-.06	.252

Afterwards, we added the T0 overconfidence variables in separate models. Each model controlled for the variables in Model 0 and was independent of all other models. The metacognitive overconfidence bias measures were predictive of fake news vulnerability in the case of verbal working memory and general knowledge (H1.2: $B = .03$, $SE = .01$, $p = .008$ and respectively H2.2: $B = .29$, $SE = .10$, $p = .004$), while metacognitive sensitivity measures were not (H1.1, H2.1, see Table 4).

Table 4. Predictive power of overconfidence in working memory and general knowledge on fake news vulnerability at T0

	Verbal Working Memory		General Knowledge	
	Metacognitive sensitivity	Metacognitive bias	Metacognitive sensitivity	Metacognitive bias

Model specific overconfidence variable	Model 1.1		Model 1.2		Model 2.1		Model 2.2	
	B	β	B	β	B	β	B	β
Constant	15.26**		15.26**		15.26**		15.26**	
Cognitive reflection	.10	.05	.09	.05	.09	.05	.08	.04
Overclaiming	.01	.01	-.01	-.01	.01	.02	.00	.00
Bullshit receptivity	.07	.09	.06	.08	.08	.09	.07	.08
Actively open-minded thinking	-.03	-.06	-.03	-.05	-.03	-.05	-.02	-.05
Model specific overconfidence variable (T0)	.00	.06	.04**	.14	.02	.08	.05**	.15
R^2	.02		.04		.02		.04	
F	1.51		2. 79**		1.90		3.06**	
ΔR^2	.003		.018		.007		.021	
ΔF	1.28		7.10**		2.75		8.42**	

Note: * $p < .05$, ** $p < .01$; Each model tests the associated hypothesis

To analyse changes in fake news vulnerability between the experimental and control groups we performed a one-way ANCOVA, having the group as an independent variable, fake news vulnerability at T1 as the dependent variable and fake news vulnerability at T0 as a covariate. The results showed that the model was not statistically significant ($F(2, 392) = 1.84$, $p = .161$, *partial* $\eta^2 = .01$), indicating that the feedback received by the two groups did not affect their assessment of the fake news articles or that the study lacked the statistical power to detect the effect of the feedback (H3.1). Regardless, the preregistered pairwise comparisons and the mediation analysis were not applicable (H3.2).

Exploratory pairwise comparisons (Table 2) indicate that the only significant between-groups differences can be seen at post-test in working memory sensitivity and bias ($t = 2.38$, $p = .018$ and $t = 2.25$, $p = .025$).

Discussion

The current study aimed to investigate the role of memory overconfidence in predicting health-related fake news vulnerability and how trust in fake news is influenced after being presented with overconfidence-correcting feedback. Measures of reflexive and reflective open-minded thinking, which have been established as predictors of political fake news (Pennycook and Rand, 2020), were included in the analysis to test whether they were also predictive of health-related fake news and whether memory overconfidence explained unique variance beyond that accounted for by these variables.

None of the reflective and reflexive open-minded thinking constructs reached statistical significance as predictors of health-related fake news vulnerability (Table 3). Given that the randomly generated pseudo-profound items were constructed starting from a database of tweets by Deepak Chopra (Pennycook et al., 2015), a proponent of holistic, alternative medicine, they share a conceptual basis with the fake news articles selected for this study and use similar language. The absence of a relationship is therefore surprising. The results showing no relationship between cognitive reflection and fake news vulnerability were also surprising given the findings of Scherer et al. (2020) and Pennycook and Rand (2020). However, Mustață et al. (2023) found a similar absence of effect in their investigation of security and defence fake news vulnerability in Central and Eastern Europe. Similar to studies of Western populations, they found an association between fake news vulnerability and actively open-minded thinking, which was not found in the current study.

Both measures of metacognitive bias were predictive of fake news vulnerability (B = .04, SE = .02, p = .008 for verbal working memory and B = .05, SE = .02, p = .004 for general knowledge). While the effect size detected was small, our preregistered analysis accounted for this possibility. However, neither metacognitive sensitivity regarding verbal working

memory nor that regarding general knowledge proved predictive of fake news vulnerability. Performance scores or summed confidence scores for correct and incorrect answers were not predictive of fake news vulnerability (see Supplementary Materials). The observed predictive power of only the metacognitive bias measures suggests that the pivotal factor isn't merely the gap between an individual's confidence and accuracy. Instead, it's a more general tendency to exhibit overconfidence, irrespective of the actual correctness of their responses. Essentially, those who don't acknowledge their errors are more susceptible to fake news, regardless of the confidence they place in their correct answers. This finding aligns with previous literature indicating that overconfidence in one's reasoning and abilities are predictive of susceptibility to misinformation (Lyons et al., 2021, Vranic et al., 2022).

While we can confidently state that overconfidence in memory predicts health-related fake news vulnerability, the specific mechanisms underlying this relationship warrant further examination. We started from the hypotheses that overconfidence in working memory will reduce responses typically seen when encountering local (i.e., intra-text) conflict, e.g., returning to previous paragraphs, and that overconfidence in general knowledge will produce fewer responses typical of detecting global conflict (i.e., between information presented in the text and previously held knowledge), e.g., verifying aspects that are uncertain. Although the current research design cannot provide the specific evidence needed to validate these hypotheses, the findings are promising as the pattern of results is consistent with this theoretical interpretation. Specifically, the fact that metacognitive bias, but not metacognitive sensitivity, predicted fake news vulnerability suggests that the critical factor may be a generalized tendency to trust one’s internal representations rather than a failure to discriminate between correct and incorrect responses. From this perspective, memory overconfidence may function by dampening the cognitive signals that normally prompt further scrutiny, thereby allowing misleading information to be accepted without challenge.

The experimental part of the current research aimed to determine whether feedback could reduce overconfidence and fake news vulnerability. After the initial tasks, the control group received feedback on the time they took to complete the working memory and general knowledge tasks. The experimental group was informed of their correct and incorrect answers and the summed confidence ratings of each. Regardless of the type of feedback received, participants' belief in fake news remained unchanged. Significant between-groups differences were observed on both measures of working memory overconfidence, with the experimental group showing less metacognitive bias and more sensitivity. This suggests that the experimental manipulation was effective in reducing working memory overconfidence.

Our intervention was similar to the one proposed by Lyons et al. (2021). However, we diverged in a key area: our feedback targeted memory overconfidence rather than overconfidence in fake news detection. Based on their model, one might predict a context-specific decrease in susceptibility to fake news; that is, when participants are made aware of the potential presence of fake articles, they're likely to approach subsequent articles with increased skepticism. However, our data do not confirm a direct causal relationship between memory overconfidence and fake news vulnerability.

Furthermore, despite evidence from previous research suggesting that accuracy nudges can reduce misinformation discernment and sharing (Pennycook et al., 2021b, Mirhoseini et al., 2023), our study did not replicate this effect. The inattention-based account of misinformation sharing suggests that different aspects of the hypercomplex social media environment control sharing behavior, irrespective of how accurately the news is perceived.

Additionally, according to the motivated system 2 reasoning (Kahan, 2016) approach, the reason people avoid revisiting certain information may not be due to a failure in detecting conflicts, but rather a reluctance to engage further with that specific information. This may

account for the changes in overconfidence, but not in fake news vulnerability, as observed in the current study.

Taken together, these findings suggest that reducing overconfidence at the level of basic cognitive operations is not sufficient, by itself, to alter how individuals evaluate complex, belief-relevant information such as fake news. Although participants became better calibrated about the limits of their memory performance, this recalibration did not generalize to the epistemic judgments required when reading news articles. One plausible interpretation is that evaluations of fake news draw on a broader constellation of cues than memory confidence alone, including identity-relevant beliefs, emotional reactions, narrative coherence, and perceived social meaning. In everyday information environments, people evaluate news content as socially and morally situated material. As a result, improvements in metacognitive accuracy may remain compartmentalized unless the feedback explicitly targets the evaluative context in which misinformation is encountered.

Our results should be interpreted with caution because of certain limitations. First, all study materials were presented in Romanian. As a result, cultural cognition, viewed as a form of motivated system 2 reasoning (Kahan, 2016, Mustață et al., 2023) and particular to post-communist countries, may have influenced our outcomes. However, studies from Ukraine, another post-communist country, have shown alignment with the existing literature (Erlich et al., 2022). This suggests that such cultural specificities may not have significantly influenced our findings. Nevertheless, the use of a convenience sample requires added caution, as it restricts the extrapolation of our findings to the broader Romanian population. Given the novelty of the aspects investigated and the experimental nature of the design, we consider the results to be relevant.

Moreover, it is important to note that our sample, which is predominantly young, female, and made up of undergraduates, is not nationally representative. According to a UK study by King and Greene (2024), being female increases vulnerability to health-related fake news, while education does not predict trust in fake news. Similarly, Arin et al. (2023) found that female participants in the UK were more vulnerable to political fake news, in contrast to their German counterparts, where education was predictive but gender was not. The demographic of well-educated, young, female university students could explain some of the differences observed in our study. These findings highlight the need for further cross-cultural studies to explore other factors that might influence these outcomes.

Another important limitation of the current research is that the design does not permit direct testing of the hypothesized mechanisms linking memory overconfidence to local and global conflict detection. The measures used relied on aggregated trust judgments for complex, full-length articles, without access to fine-grained indicators of how participants interacted with specific text segments. As a result, it is not possible to determine whether overconfident individuals failed to detect conflicts, detected them but chose not to act on them, or engaged in alternative processing strategies altogether.

Future research should address this limitation by incorporating process-level methodologies. One strategy would be to employ eye-tracking software to observe whether participants exhibiting working memory overconfidence will display fewer back-and-forth movements, suggesting a lack of local conflict detection. Another strategy would be to provide a search function while reading the articles and inform participants that they can search for information about which they are unsure. In such a design, we would expect that those with increased general knowledge overconfidence would be less likely to use the search function. These designs can address another limitation of the current study, namely the articles used to measure fake news vulnerability are complex text and we relied on general, overarching

assessments, without knowing how participants related to the different paragraphs. Analysing the contextual factors that trigger (dis)trust (e.g., skipping to the end of the article) could further advance our understanding of fake news and how to combat its influence.

While we expected that our feedback procedure will act as an accuracy nudge, it is possible that the feedback given was not strong enough to transfer the reduction in overconfidence from memory tasks to the context of fake news articles. The previously suggested eye-tracking approach could provide molecular evidence regarding the extent of feedback transfer between tasks. This would be particularly informative if paired with different types of feedback, including feedback directly related to the news articles. Future studies should explore whether feedback that corrects overconfidence can serve as an effective accuracy nudge in social media environments.

Finally, a comprehensive examination aimed to differentiate between intuitive and motivate reasoning would require to include measures of belief in complementary and alternative medicine and physiological responses to various text segments (e.g., using the previously discussed methods), complemented by qualitative research on participants' existing beliefs about the content. Together, these methods would provide nuanced evidence for the debate between the classical reasoning model and motivated system 2 reasoning.

The current study found that memory overconfidence is a robust, experimentally mutable (i.e., changeable) predictor of fake news vulnerability when assessed alongside established predictors, opening new fake news research directions. Causal relationships and underlying mechanisms need to be further explored in different contexts.

Study 3. Impact of Repeated Exposure to Polarized Health-Related News on Attitudes Toward Dietary Supplements

The aim of this paper is to investigate how repeated exposure to polarized health information influences explicit and implicit attitudes over the course of two weeks. Prior studies on the illusory truth effect and evaluative conditioning have typically relied on isolated statements or short stimuli, which limits ecological validity. To better approximate real-world information exposure, the current study is, to our knowledge, the first to use full-length online articles to simulate the experience of encountering health-related content in a digital news environment.

Participants were randomly assigned to one of four exposure conditions differing in informational valence: a PRO group (exposed only to favorable articles about dietary supplements), a CON group (exposed only to unfavorable articles), a MIX group (exposed to both positive and negative articles), and a Control group (exposed to neutral, space-related content). After group assignment and the initial assessment, participants received one article per day corresponding to their assigned condition. Explicit and implicit attitudes toward dietary supplements were assessed at three time points: baseline (T0), one week (T1), and two weeks (T2), allowing for the examination of both within-subject attitude change and between-group differences following repeated exposure.

Hypotheses

(H1a) The PRO group will have more favorable implicit attitudes towards dietary supplements in the final assessment than it did at the initial assessment.

(H1b) The PRO group will have more favorable explicit attitudes towards dietary supplements in the final assessment than it did at the initial assessment.

(H2a) The CON group will have less favorable implicit attitudes towards dietary supplements in the final assessment than it did at the initial assessment.

(H2b) The CON group will have less favorable explicit attitudes towards dietary supplements in the final assessment than it did at the initial assessment.

(H3a) Implicit attitudes in the MIX group will be different at the final assessment as compared to those in the initial assessment.

(H3b) Explicit attitudes in the MIX group will be different at the final assessment as compared to those in the initial assessment.

(H4a) The PRO group will have more favorable implicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H4b) The PRO group will have more favorable explicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H5a) The CON group will have more favorable implicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H5b) The CON group will have more favorable explicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H6a) The MIX group will have more favorable implicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H6b) The MIX group will have more favorable explicit attitudes towards dietary supplements in the final assessment when comparted to the control group.

(H7a) The relationship between group and implicit attitude change is moderated by initial attitudes towards dietary supplements; More positive initial attitude and will lead to less pronounced attitude changes in the PRO group and more pronounced attitude changes in the CON group.

(H7b) The relationship between group and explicit attitude change is moderated by initial attitudes towards dietary supplements; More positive initial attitude and will lead to less pronounced attitude changes in the PRO group and more pronounced attitude changes in the CON group.

(H8a) The relationship between group and implicit attitude change is mediated by the time spent reading the news articles.

(H8b) The relationship between group and explicit attitude change is mediated by the time spent reading the news articles.

(H9a) The moderating effect of initial attitudes and similar news exposure on the relationship between group and implicit attitude change is mediated by the time spent reading the news articles.

(H9b) The moderating effect of initial attitudes and similar news exposure on the relationship between group and explicit attitude change is mediated by the time spent reading the news articles.

Repeated exposure to congruent information was expected to strengthen attitudes in the direction of the presented valence. Specifically, participants repeatedly exposed to positive content (PRO group) were expected to develop significantly more favorable implicit (H1a) and explicit (H1b) attitudes toward dietary supplements at the final assessment compared to baseline, while those repeatedly exposed to negative content (CON group) were expected to show less favorable attitudes (H2a, H2b). Participants who encountered both positive and negative information (MIX group) were expected to also show differential changes across time (H3a, H3b). Comparisons with the Control group were expected to confirm that attitude shifts were driven by informational valence rather than mere passage of time or task

repetition (H4a, H4b, H5a, H5b, H6a, H6b for PRO-implicit, PRO-explicit, CON-implicit, CON-explicit, MIX-implicit and MIX-explicit respectively).

Beyond these primary effects, the study also examined how initial attitudes shaped evaluative change. It was anticipated that participants holding more positive initial views of dietary supplements would show smaller changes in the PRO condition but larger negative shifts in the CON condition (H7a, H7b), based on how diagnostic the nature of the presented information is in regards to previously held beliefs. Moreover, the relationship between informational valence and attitude change was expected to be mediated by engagement, such that greater time spent reading articles would predict stronger attitude change (H8a, H8b). Finally, this mediation was hypothesized to be moderated by initial attitudes, reflecting the interaction between prior beliefs and the effects of repeated exposure (H9a, H9b).

Method

The preregistration for the study can be accessed at https://osf.io/r7y8m (see description). Open Science Framework data and code can be accessed at: https://osf.io/p6xw5

Design and Procedure

The study employed a 4 × 3 mixed experimental design. The between-subjects factor was Group, defined by the type of articles participants read (PRO - positive information about dietary supplements; CON - negative information about dietary supplements; MIX - both positive and negative information regarding dietary supplements; or Control - neutral content about space exploration, unrelated to health), and the within-subjects factor was Time, with assessments at baseline (T0), one week (T1), and two weeks (T2). Participants were randomized with replacement by the Gorilla online experimental platform (Anwyl-Irvine et al., 2020) into one of the four groups before exposure began.

Articles were sourced from diverse online outlets and standardized for length, structure, and believability. When an original article contained both positive and negative information, it was divided into two separate stimuli to maintain experimental control over valence.

Participants accessed the study via a link to the Gorilla platform. After providing informed consent, they were presented with a cover story stating that the purpose of the project was to assess personal characteristics of readers who enjoy different types of online articles. Following the initial assessment (T0), participants were randomized into one of the four groups and presented with the first article of their assigned condition. After reading, they rated the article on a five-star scale and were informed that their next article would be available the following day.

Each participant then received one article per day for six consecutive days, followed on the seventh day by the second assessment (T1). The procedure was repeated for a second cycle of six daily articles, after which participants completed the final assessment (T2). The Gorilla platform automatically recorded time spent reading each article (in seconds). Each assessment included measures of explicit attitudes, implicit attitudes, and frequency of online exposure to health information that either aligned with or contradicted the study content.

Due to a technical error in the platform, participants in the MIX group did not receive articles in the intended randomized order. Instead, they were systematically presented with a positive article followed by a negative one, such that the day immediately preceding each assessment (T1 and T2) always featured a negative (CON-type) article. This deviation was consistent across all MIX participants and was documented prior to data analysis.

Participants

Participants were recruited from the Babeș-Bolyai University student population, who were offered course credit in exchange for participation, and from the general Romanian

population through social media posts and targeted online advertisements. Inclusion criteria required participants to be at least 18 years old and to have access to a keyboard-compatible device for completing the online tasks.

A power analysis based on the procedures described by Preacher et al. (2007) indicated that a total sample of 200 participants (50 per group) would provide sufficient power to detect a medium-sized conditional indirect effect in moderated mediation models. Recruitment was planned to reach this target, and data collection concluded once it was achieved, with participants who had already begun the study allowed to complete their participation.

A total of 345 participants initially enrolled in the study. Of these, 117 were excluded for not completing the second assessment, resulting in a final sample of 228 participants (174 women and 52 men, M_{age} = 20.81 years, SD_{age} = 4.88). In terms of education, 1 participant reported completing primary education, 169 secondary education, 42 undergraduate studies, and 16 graduate studies. Participants were distributed across conditions as follows: PRO (n = 68), CON (n = 51), MIX (n = 52), and Control (n = 57).

Measures

a. Implicit Attitudes

Implicit attitudes toward dietary supplements were assessed using a computerized Implicit Association Test (IAT; Greenwald et al., 1998) administered via the Gorilla online platform. The IAT measured the relative strength of automatic associations between dietary supplements and positive versus negative evaluative attributes.

Participants categorized stimuli into one of two target categories: dietary supplements (e.g., “vitamins”, “minerals”, “nutrients”) and objects (e.g., “trousers”, “chair”, “poster”), and two attribute categories: positive (e.g., “healthy”, “good”, “safe”) and negative (e.g., “risky”, “bad”, “dangerous”). During the task, single words appeared sequentially in the center of the

screen and participants were instructed to classify each as quickly and accurately as possible using two response keys corresponding to the category pairings displayed on-screen. The task alternated between congruent blocks (e.g., dietary supplements + positive / objects + negative) and incongruent blocks (e.g., dietary supplements + negative / objects + positive). Reaction times and accuracy were recorded. Incorrect responses triggered an error message, and the trial continued only after the correct key was pressed, penalizing errors through longer response times (Greenwald et al., 2003).

An IAT D-score was computed for each participant at each time point (T0, T1, T2) following standard scoring procedures (Greenwald et al., 2003). The D-score was calculated by subtracting the mean reaction time for congruent blocks from that for incongruent blocks and dividing by the pooled standard deviation of all response latencies. Higher D-scores indicated more favorable implicit attitudes toward dietary supplements.

b. Explicit attitudes

Explicit attitudes were measured using three visual analogue scale (VAS) questions: “To what degree do you consider dietary supplements to be efficient?”, “To what degree do you consider dietary supplements to be harmful?” (reverse-scored), “To what degree would you recommend dietary supplements to a loved one?”. Scores were aggregated across items at each time point to yield an explicit attitude index, with higher values reflecting more favorable evaluations. Internal consistency was acceptable at every evaluation (Cronbach’s α = .79, .75, and .76 respectively).

The frequency of exposure to health-related news and similar news was assessed using the following questions: “How often do you read health-related news?” (only at T0); “During the last week, how often have you encountered positive news articles related to dietary supplements?” (only at T0), “During the last week, how often have you encountered negative

news articles related to dietary supplements?" (only at T0) "During the last week, how often have you encountered similar news articles to those presented on this platform?" (T1, T2), "During the last week, how often have you encountered news articles that contradicted those presented on this platform?" (T1, T2); "Approximately how much time did you spend last week reading health-related news?".

Statistical Analysis

Data were analyzed using mixed ANCOVA models with Group as the between-subjects factor and Time as the within-subjects factor. Dependent variables were explicit and implicit attitude scores. Covariates included time spent reading health-related news and self-reported exposure to health-related news on the topic of dietary supplements, assessing both similar and contradictory.

Planned contrasts compared T2-T0 changes within the PRO, CON, and MIX groups and between-group differences at T2. One-tailed significance tests were used in accordance with our directional hypotheses (PRO and CON), and two-tailed tests for the nondirectional hypothesis about the MIX group.

Moderation, mediation and moderated mediation analyses were conducted using PROCESS (Hayes, 2017). These analyses tested whether baseline attitudes moderated the effect of group assignment on attitude change, whether time spent reading mediated this effect, and whether this mediation was further moderated by initial attitudes or exposure to similar news. Indirect effects were evaluated using bias-corrected bootstrapped confidence intervals (BootCI).

All data was analysed using IBM SPSS 25.

Results

Descriptive statistics for the primary study variables are presented in Table 1. These include explicit and implicit attitudes toward dietary supplements, assessed at baseline (T0), one week (T1), and two weeks (T2), as well as a measure of participant engagement with the study materials, operationalized as the average time spent reading the articles.

Self-reported frequency of exposure to health-related news, time spent reading such news, and exposure to positive and negative articles about dietary supplements outside the experimental context were also assessed at each time point and included in the analyses as control variables. Full descriptive data for these control measures are available in the Supplementary Materials (Table S1).

Table 1. Descriptive statistics

	PRO (n = 68)		CON (n = 51)		MIX (n = 52)		Control (n = 57)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
T0								
Implicit Attitudes	-0.08	0.42	0.00	0.48	-0.12	0.46	-0.17	0.43
Explicit Attitudes	190.21	62.74	202.25	56.37	200.17	44.44	203.60	48.83
Efficiency rating	61.99	24.68	66.86	20.82	67.79	16.83	67.18	20.17
Harmfulness rating	29.00	19.87	24.25	18.10	29.06	20.62	26.16	17.44
Recommendation willingness	56.22	27.11	58.65	25.43	60.44	19.42	61.58	22.79
Time spent reading study articles (s)	1604.25	6854.92	1133.38	3737.68	1269.83	3982.75	878.71	2289.01
T1								
Implicit Attitudes	-0.05	0.43	-0.13	0.47	-0.30	0.47	-0.18	0.45

Explicit Attitudes	201.31	51.03	150.80	57.67	184.92	41.35	188.12	49.91
Efficiency rating	67.93	21.33	49.82	21.43	64.29	18.69	62.72	20.07
Harmfulness rating	27.69	20.65	45.35	22.19	35.02	18.06	28.77	18.48
Recommendation willingness	60.07	22.32	45.33	23.12	54.65	21.34	53.18	23.06
T2								
Implicit Attitudes	-0.15	0.41	-0.09	0.39	-0.23	0.41	-0.13	0.40
Explicit Attitudes	206.15	49.22	148.06	55.75	176.40	45.61	187.54	52.23
Efficiency rating	68.94	18.67	47.37	22.13	61.58	20.45	60.65	22.06
Harmfulness rating	26.01	18.36	44.08	20.93	37.62	21.13	25.98	16.56
Recommendation willingness	62.22	20.84	43.76	23.65	51.44	23.20	51.88	24.04

To analyze changes in implicit attitudes a 4x3 mixed-design ANCOVA was conducted on the IAT D-scores at the three assessments, having the group as an independent variable. Time spent reading health-related news, frequency of exposure to health-related information (T0, T1, T2), and time spent reading the articles pro and against dietary supplements (T0, T1, T2) were included as covariate. Mauchly's test indicated that the assumption of sphericity was met (p = .856). The analysis indicated that the interaction effect between group and time is not statistically significant ($F(6, 428) = 1.90$, p = .080, η^2_p = .03),

To examine changes in implicit attitudes, a 4 × 3 mixed-design ANCOVA was conducted on IAT *D*-scores across the three assessments, with group (PRO, CON, MIX, Control) as the between-subjects factor and time (T0, T1, T2) as the within-subjects factor. Time spent reading health-related news, frequency of exposure to health-related information (T0, T1, T2), and time spent reading pro- and anti-supplement articles (T0, T1, T2) were included as covariates. Mauchly's test indicated that the assumption of sphericity was met, p = .856. The analysis revealed that the Group × Time interaction was not statistically significant, $F(6, 428) = 1.90$, p = .080, η^2_p = .03. Given the absence of a significant interaction, no follow-up

contrasts were performed. Hence, the collected data was insufficient to support H1a, H2a, H3a, H4a, H5a, H6a.

To analyze changes in explicit attitudes toward dietary supplements, a 4 × 3 mixed-design ANCOVA was conducted on the composite explicit attitude scores, with group (PRO, CON, MIX, Control) as the between-subjects factor and time (T0, T1, T2) as the within-subjects factor. Time spent reading health-related news, frequency of exposure to health-related information (T0, T1, T2), and time spent reading pro- and anti-supplement articles (T0, T1, T2) were included as covariates.

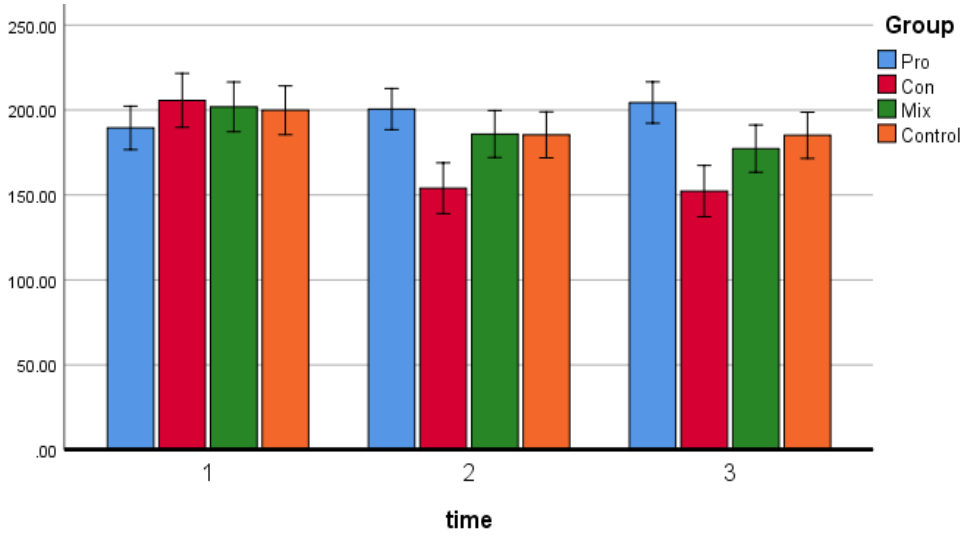
Mauchly's test indicated that the assumption of sphericity was violated, $W = .809$, $\chi^2(2) = 45.05$, $p < .001$; therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = .84$). The analysis revealed a significant Group × Time interaction, $F(5.04, 428) = 14.34$, $p < .001$, $\eta^2_p = .167$, showing that changes in explicit attitudes differed across exposure conditions.

Pre-registered within-group comparisons showed that participants in the PRO condition developed significantly more favorable explicit attitudes from baseline to the final assessment ($M\Delta = -14.86$, $SE = 5.93$, $p_{1-tailed} = .019$, $d = .30$). Participants in the CON condition showed the opposite pattern, reporting less favorable attitudes over time ($M_\Delta = -53.42$, $SE = 7.35$, $p_{1-tailed} < .001$, $d = 1.02$). The MIX group exhibited a significant change consistent with the most recent exposure (i.e., negative article, $M_\Delta = -24.56$, $SE = 6.77$, $p = .001$, $d = .50$), whereas the Control group showed no significant change ($M_\Delta = -14.74$, $SE = 6.63$, $p = .080$, $d = .29$). H1b, H2b, H3b are therefore supported by the collected data.

Between-group comparisons at the final assessment (T2) confirmed this pattern. The CON group reported significantly lower explicit attitudes than the Control group ($M_\Delta = -32.84$, $SE = 10.75$, $p = .015$, $d = .32$). Explicit attitudes in the PRO group were not significantly

different from the control group (M_{Δ} = 19.24, *SE* = 9.29, *p* = .215, *d* = .20) but were significantly higher than in the CON group (M_{Δ} = 52.08, *SE* = 10.03, *p* < .001, *d* = .52), and than in the MIX group (M_{Δ} = 27.07, *SE* = 9.38, *p* = .026, *d* = .29). No other between-group differences were significant. Data provided clear support for H5b and partial support for H4b. Together, these results indicate that repeated exposure to polarized information altered explicit, but not implicit, attitudes toward dietary supplements, and that these changes occurred in the direction of the informational valence of the exposure (see Figure 1).

Figure 1. *Changes in explicit attitudes toward dietary supplements across time as a function of exposure condition.*



In order to test whether baseline attitudes moderated the relationship between group assignment and attitude change, two multiple regression model were conducted. The model having implicit attitude change as criterion was statistically significant ($F(11, 216) = 16.80$, $R^2 = .46$, $p < .001$), but the only significant predictor was baseline implicit attitudes ($b = -.84$, $SE = .11$, $p < .001$). For the model having explicit attitude change as criterion, neither implicit ($F(3, 216) = .04$, $\Delta R^2 = .00$, $p = .989$) nor explicit attitudes ($F(3, 216) = 1.94$, $\Delta R^2 = .02$, $p = .124$) moderated the effect. Thus, the collected data was insufficient to support H7a and H7b.

Given that the mixed-design ANCOVA revealed no significant interaction between time and group for implicit attitudes and no consistent within-group changes, mediation analyses involving implicit attitudes as the outcome were not pursued. Therefore, H8a and H9a are not supported.

For explicit attitudes, the simple mediation model testing whether time spent reading mediated the relationship between exposure condition and explicit attitudes at T2 failed to meet the assumptions for mediation, as the path from Group to Reading time was nonsignificant ($F(3, 224) = .26$, $p = .856$, $R^2 = .003$). The collected data was not sufficient to support H8b.

To further examine whether engagement effects depended on participants' initial evaluations, two preregistered moderated mediation models (PROCESS Model 7; Hayes, 2022) were tested. When baseline implicit attitudes were included as a moderator of the Group to Reading (a) path, the overall model was still not significant ($F(7, 220) = 1.48$, $p = .175$, $R^2 = .045$), although the interaction effect was significant ($F(3, 220) = 3.01$, $p = .031$, $\Delta R^2 = .039$). In accordance, no conditional indirect effects were significant, all indices of moderated mediation included zero (PRO = -2.59, 95% CI [-7.41, 0.87]; CON = -3.16, 95% CI [-7.77, 1.26]; MIX = 3.83, 95% CI [-9.62, 1.43]). The analysis with baseline explicit attitudes as the moderator also yielded a nonsignificant model ($F(7, 220) = 0.94$, $p = .475$, $R^2 = .03$). Therefore, H8b is not supported by the collected data.

To further examine the changes in explicit attitudes, the three items composing the composite score (i.e., perceived efficiency, perceived harmfulness, and recommendability of dietary supplements) were analyzed separately using 4 × 3 mixed-design ANCOVAs with the same covariates as the main model. These analyses were treated as exploratory as they have not been preregistered.

The multivariate test revealed a significant time × group interaction, $F(18, 1344) = 6.27$, $p < .001$, $\eta^2_p = .078$. At the univariate level, all three attitude items showed significant time × group interactions (all *ps* < .001, *partial* η^2s = .09-.14).

Perceived efficiency increased significantly in the PRO group from T0 to T2 ($M_\Delta = 6.96$, $p = .015$, $d = .42$) and decreased significantly in the CON group ($M_\Delta = -19.49$, $p < .001$, $d = .83$). At T2, participants in the CON group rated supplements as less efficient than those in the other groups (*ps* ≤ .006, *ds* = .64-1.04), which did not differ from each other.

Harmfulness increased significantly in the CON group from T0 to T2 ($M_\Delta = 19.82$, $p < .001$, $d = .84$) and to a smaller extent in the MIX group ($M_\Delta = 8.56$, $p = .004$, $d = .48$). At T2, these groups judged supplements as more harmful than both the PRO (*ps* ≤ .007, *ds* = .60 and .94) and Control groups (*ps* ≤ .011, *ds* = .61 and .94).

Recommendability decreased significantly in the CON group from T0 to T2 ($M_\Delta = -14.88$, $p < .001$, $d = .77$) and to a lesser degree in the MIX group ($M_\Delta = -9.00$, $p = .008$, $d = .52$). At T2, the CON group reported lower recommendability than the PRO group ($p < .001$, $d = .82$).

Discussions and conclusions

The present study investigated how repeated exposure to polarized health-related content influences explicit and implicit attitudes toward dietary supplements. Over two weeks of daily exposure to full-length articles, participants' explicit attitudes shifted in the direction of the presented valence (H1b, H2b, H3b, partial H4b, H5b), while implicit attitudes remained stable. The expected moderation by initial attitudes (H7), mediation through engagement (i.e., time spent reading, H8) and the hypothesized moderation by initial attitudes of the mediated effect (moderated mediation, H9) were not supported. Exploratory analyses of the three explicit attitude components (i.e., perceived efficiency, harmfulness, and recommendability)

showed distinct valence-dependent changes, indicating that repetition differentially shaped specific evaluative dimensions rather than producing uniform shifts.

Consistent with predictions derived from the illusory truth effect (Hasher et al., 1977; Fazio et al., 2019) and the mere exposure effect (Zajonc, 1968; Bornstein and Craver-Lemley, 2022), repeated exposure to polarized health-related content systematically modified explicit evaluations. Participants repeatedly exposed to favorable articles developed more positive explicit attitudes toward dietary supplements (H1b), while those exposed to negative articles developed more negative attitudes (H2b). The Control group, who read unrelated content, showed no change, confirming that attitude shifts were due to informational valence rather than time or repeated testing. Repeated reading of congruent information, even when spread across distinct articles, appears sufficient to increase the perceived validity and desirability of the presented perspective.

These results align with prior findings that repetition enhances processing fluency, which is then misattributed to truth (Unkelbach and Greifeneder, 2013). Because all articles were matched in length, structure, and evidential strength, the observed changes in explicit attitudes are unlikely to reflect persuasion via argument quality and instead point to fluency-driven propositional updating. The pattern observed in the mixed-exposure group provides further support for this interpretation. Participants in this condition consistently displayed negative shifts in their explicit attitudes (H3b), which due to a technical ordering error was the most recently encountered valence before each assessment point, suggesting that in contexts of informational ambivalence, recency effects can outweigh the cumulative balance of prior exposures. This is consistent with models of sequential belief updating showing that when conflicting information is presented sequentially and is comparable in evidentiary strength, recency exerts a disproportionate influence on judgments (Hogarth and Einhorn, 1992). Nonetheless, because negative information is oriented towards more often and

weighted more strongly than positive information (the “bad is stronger than good” phenomenon; Baumeister et al., 2001), the present results cannot rule out the possibility that negativity dominance, rather than recency alone, contributed to the observed pattern.

In practical terms, this suggests that individuals may come to view dietary supplements as more effective, less harmful, or more recommendable simply because related claims are encountered repeatedly in everyday information streams. Importantly, the absence of change in implicit attitudes indicates that these shifts remain largely declarative and context-dependent rather than fully internalized. Explicit attitudes appear flexible and sensitive to recent informational input, while underlying associative structures remain stable over short exposure windows. This dissociation helps explain why misinformation can rapidly influence stated beliefs and intentions without necessarily producing durable changes in automatic responses, and why corrective efforts may succeed at the level of explicit judgment while failing to alter more pervasive evaluative tendencies. In algorithmically curated environments, where repetition and recency are structurally amplified, such surface-level but behaviorally relevant attitude shifts may nevertheless accumulate into meaningful public health consequences, particularly when decisions rely on explicit reasoning rather than implicit preferences.

The exploratory analyses revealed that the CON group exhibited the strongest and most consistent shifts across all three evaluative dimensions: efficiency, harmfulness, and recommendability. The MIX group showed smaller but still significant increases in perceived harmfulness and decreases in recommendability, which suggests that repetition produces cumulative effects even when exposures are heterogeneous. These changes may reflect partial counterbalancing from the positive articles, the recency of the last negative exposure, or both. Because both harmfulness and recommendability displayed a consistent trend from T0 to T1 to T2 (with a significant change from T0 to T1 only for harmfulness, $M_{\Delta} = -6.49$, $p = .043$),

the data suggests an additive negative exposure effect that was tempered, though not eliminated, by intermittent positive information. By contrast, efficiency increased modestly in the PRO group but was only slightly higher than in the Control group at T2 (H4b, $M_{\Delta} = 8.76$, $p = .124$). This suggests that positive claims about supplement benefits may be less impactful than warnings about potential harm. This asymmetry aligns with extensive evidence showing that negative information typically produces stronger and more persistent evaluative shifts than positive information (Soroka et al., 2019, Watson et al, 2024). Practically, these findings suggest that health related judgments, especially those concerning perceived harm and social recommendation, are highly susceptible to repeated negative framing, whereas positive framing appears weaker and more fragile, but nevertheless offers some degree of protection when mixed in an environment saturated with negative exposure.

The observed effects extend prior findings based on brief or decontextualized statements (Unkelbach and Speckmann, 2021; Morgan and Cappella, 2023), showing that fluency-driven evaluative shifts are not limited to aphoristic or slogan-like content. Instead, similar shifts arise from repeated exposure to full-length, naturalistic articles, suggesting that real-world news formats, despite their complexity and narrative structure, can still produce the same fundamental repetition-based distortions in explicit evaluations.

As online news consumption is driven primarily by negativity (Robertson et al., 2023), health-related content may be especially vulnerable to this dynamic due to its high personal relevance and the emotional weight carried by risk-related information. When algorithms amplify engagement signals, users are more likely to encounter repeated negatively framed health content, which our findings suggest can disproportionately influence evaluations of harmfulness and recommendability. In real-world settings where user interests and algorithmic curation jointly shape exposure, this may create a self-reinforcing loop of negative information. Practically, this underscores the value of deliberately incorporating

balanced or positively framed health information. Further research on recency and timing is required to determine whether such content should be introduced strategically before important health decisions, in order to counteract the cumulative impact of negativity-heavy feeds and support more proportionate and evidence-aligned judgments.

Taken together, this pattern highlights an important asymmetry in how evaluative dimensions respond to repeated exposure. Judgments related to risk and social endorsement appear especially sensitive to cumulative negative information, likely because they map onto precautionary decision-making processes where avoiding harm is prioritized over maximizing potential benefits. From a behavioral standpoint, this means that even modest increases in perceived harmfulness, when repeated, may disproportionately reduce willingness to recommend or adopt health-related products, independently of beliefs about their effectiveness. In contrast, efficiency-related beliefs seem more resistant to change and require stronger or more consistent positive input to shift meaningfully. In real-world contexts, such as decisions about supplement use, this imbalance suggests that individuals may continue to acknowledge potential benefits while simultaneously disengaging behaviorally due to heightened concern about risks. Consequently, misinformation ecosystems dominated by negative framing may suppress health behaviors not by convincing individuals that interventions are ineffective, but by amplifying uncertainty, caution, and social reluctance. This distinction is relevant for designing corrective strategies, as countering exaggerated harm perceptions may require sustained, well-timed, and coherent positive information rather than isolated factual rebuttals.

In contrast to the clear shifts observed in explicit evaluations, implicit attitudes toward dietary supplements remained unchanged across the two-week exposure period (H1-6a). This pattern aligns with contemporary evidence showing that changes in implicit evaluations depend strongly on the type, diagnosticity, and consistency of new information, the strength and age

of prior attitudes, and the timescale of exposure (Kurdi and Charlesworth, 2023). Implicit attitudes are most likely to shift when the target is novel or weakly evaluated, when new information is strongly diagnostic or affectively potent, or when repeated pairings occur under tightly controlled conditions.

The present attitude object, dietary supplements, is familiar, frequently discussed, and carries a predominantly positive prior valence for our target population (Burcă et al., 2022). Implicit attitudes toward well-known targets tend to show reduced malleability and may be buffered by entrenched associative structures (Van Dessel et al., 2017). Additionally, because the articles were balanced in structure and intentionally non-sensational, the affective input may have been insufficient to trigger strong evaluative conditioning effects (McConnell and Rydell, 2014). Repeated article exposure may therefore have lacked the associative shock or diagnostic weight needed to update established evaluations (Cone and Ferguson, 2015).

Another possibility is methodological: the IAT may be less sensitive to short-term changes when the target category is broad, multifaceted, or already strongly encoded (Van Dessel et al., 2017). Tasks such as the Affect Misattribution Procedure (AMP, Payne et al., 2005) or Evaluative Priming (EPT, Spruyt et al., 2009) can detect subtler shifts in implicit evaluations of familiar objects and may therefore be better suited for future work on health-related content.

Taken together, these findings suggest that while explicit evaluations of dietary supplements readily shift under repeated exposure to polarized information, implicit attitudes may require either more potent affective signals, more diagnostic or surprising information, or longer and more diverse exposure histories to exhibit measurable change. In real-world settings, such changes may arise only when repetition is accompanied by credible social endorsement, identity relevance, or experiential consequences. Dietary supplements, being familiar and

frequently discussed in public discourse, may therefore represent an attitude domain in which everyday news exposure alone is unlikely to yield measurable changes in automatic evaluations.

Contrary to predictions, the hypothesized mediation by engagement, operationalized as time spent reading the assigned articles, was not supported (H8). Participants did not differentially invest time depending on the informational valence of the content. Similarly, the preregistered moderated mediation models testing whether baseline implicit or explicit attitudes moderated the relationship between group specific exposure and engagement were nonsignificant (H9). This indicates that prior beliefs did not systematically shape how long participants engaged with the material. Our expectation that counterattitudinal information would be perceived as more diagnostic, thereby eliciting deeper engagement and stronger propositional updating, was therefore not supported (H7). The lack of mediation by engagement is consistent with findings that exposure alone, not depth of processing, is sufficient to shift explicit evaluations, because illusory truth effects operate primarily through familiarity and fluency rather than comprehension or elaboration (Fazio et al., 2019; Unkelbach and Greifeneder, 2013). However, our engagement measure was a coarse proxy, capturing only total reading time. More fine-grained measures such as eye-tracking, attention-based metrics (e.g., attention allocation), or comprehension tests may better differentiate between superficial and elaborative processing in future work and offer clearer perspectives on their interaction with the illusory truth effect.

Limitations and further directions

Although several methodological limitations and directions for further research have already been noted in interpreting the findings, a few broader limitations and directions warrant explicit consideration.

The present study sought to improve ecological validity in the investigation of repeated exposure by using full-length, naturalistic articles rather than brief statements or isolated claims. Although this approach provides a closer approximation to real-world health communication, the exposure environment still differed from genuine online news ecosystems. Participants read a single text-only article per day whereas everyday exposure is rapid, interleaved, multimodal, and often shaped by social cues such as likes, comments, or peer sharing. Future research should incorporate more dynamic presentations, such as simulated algorithmic feeds, social endorsement signals, or mixed-content timelines, in order to capture the attentional competition and contextual layering characteristic of actual digital media environments.

A second limitation concerns the MIX condition. Although it was designed to provide balanced exposure, a technical error produced a fixed sequence in which negative content always immediately preceded each assessment. This systematic recency constraint limits our ability to interpret whether attitude trajectories reflected cumulative exposure or recency-driven updating. Fully counterbalanced designs and stimuli that integrate both positive and negative information within the same article would allow future studies to disentangle additive effects from recency, and negativity biases more precisely.

A significant limitation of this study concerns the potential for measurement insensitivity regarding implicit attitudes toward dietary supplements. As noted in the discussion, while explicit evaluations shifted, the stability of implicit scores may be an artifact of the Implicit Association Test (IAT) itself rather than a true absence of cognitive change. Because dietary supplements are deeply entrenched, positively valenced objects for this population (Burcă et al., 2022), the IAT may have lacked the sensitivity to detect nuanced, short-term updates within such broad associative structures over a two-week period. Furthermore, the stimulus characteristics, i..e, the balanced and non-sensational nature of the articles, were designed for

ecological validity but may have lacked the affective potency required to trigger measurable evaluative conditioning. Future research should address these constraints by diversifying implicit measures (e.g., employing the Affect Misattribution Procedure or Evaluative Priming Tasks), which may be more sensitive to subtle shifts in familiar health-related attitudes, testing whether more diagnostic or affectively charged information (as suggested by Cone & Ferguson, 2015) is necessary to bypass the buffer of prior entrenched beliefs, and increasing the exposure period beyond two weeks to determine if implicit structures simply require a longer timescale to align with explicit shifts.

Another limitation of the current study is the granularity of the engagement measure. It is possible that our use of total reading time as a proxy for engagement was too coarse to capture the psychological nuances of the illusory truth effect. As research suggests that familiarity and fluency often override deep comprehension (Fazio et al., 2019), our measure may have failed to distinguish between participants who merely skimmed the articles and those who engaged in the elaborative processing necessary for propositional updating. To address these measurement constraints, future research should move beyond simple time-based metrics to include objective attention metrics (e.g., utilizing eye-tracking to measure specific gaze duration on counterattitudinal arguments versus pro-attitudinal content, providing a clearer picture of attention allocation), and cognitive depth assessments (e.g., implementing comprehension tests post-exposure to empirically differentiate between superficial fluency and active elaboration).

Finally, the sample consisted primarily of university students, which restricts generalizability. Young adults differ from older or more medically vulnerable populations in terms of prior familiarity with dietary supplements (Burcă et al., 2022), susceptibility to misinformation (Ma et al., 2025), and online consumption patterns. Future work should recruit more diverse

samples, including older adults, individuals with chronic health conditions, and frequent consumers of health-related content on social media platforms.

6. General conclusions and discussion

The thesis addressed its theoretical, methodological, and practical objectives through a coordinated three-study empirical program examining health-related fake news vulnerability from cognitive, metacognitive, and environmental perspectives.

Theoretical objectives

A primary theoretical objective of the thesis was to clarify the cognitive and environmental mechanisms underlying fake news vulnerability.

The thesis first examined how cognitive style, apophenic tendencies, and news media literacy contribute to health-related fake news vulnerability. Consistent with prior literature, ontological confusion and receptivity to pseudo-profound bullshit were associated with increased vulnerability to misleading health-related content, suggesting that susceptibility partly reflects broader interpretative styles characterized by reduced sensitivity to semantic, causal, and ontological inconsistencies. At the same time, the interaction between cognitive reflection and news media literacy suggested that analytic reasoning is most protective when individuals lack structural knowledge regarding how credible journalism is typically organized and communicated. These findings support a layered model of epistemic evaluation in which personality dispositions, learned world models, and reflective reasoning jointly influence whether misleading information is experienced as coherent, suspicious, or worthy of additional scrutiny.

A second theoretical objective concerned the role of metacognitive monitoring, particularly overconfidence. The findings demonstrated that overconfidence in memory processes predicts fake news vulnerability independently of analytic reasoning and response tendencies.

The results support the interpretation that individuals who systematically overestimate the reliability of their own cognitive outputs are less likely to revisit information, detect inconsistencies, or engage in corrective scrutiny. From this perspective, vulnerability to fake news partly reflects failures in epistemic self-monitoring rather than deficits in reasoning competence alone.

The thesis also investigated how repeated exposure shapes evaluative processes in ecologically realistic informational environments. Repeated exposure to coherent narratives systematically altered explicit evaluations, even when participants recognized certain claims as false, illustrating the influence of familiarity, fluency, and continued influence mechanisms. At the same time, implicit attitudes remained comparatively stable, suggesting that repeated exposure primarily affects reflective evaluations over moderate timeframes. Together, these findings support a layered account of misinformation effects in which environmental repetition first alters explicit judgments and only later, potentially under more intensive or affectively charged conditions, may reshape automatic evaluative processes.

Taken together, the theoretical findings converge on the view that fake news vulnerability emerges from dynamic interactions between cognitive style, personality factors, metacognitive monitoring, memory-based fluency effects, and the structural properties of digital informational environments.

Methodological objectives

A central methodological objective of the thesis was to improve the ecological validity of misinformation research. To address the limitations of headline-based paradigms, the thesis developed and validated full-length misinformation and disinformation articles modeled after realistic online health-related content. Exploratory and confirmatory factor analyses demonstrated that these materials function as psychometrically valid stimuli capable of

distinguishing fake news vulnerability from trust in reliable news. This approach allowed the assessment of cognitive processes that cannot be adequately captured through isolated headlines, including narrative integration, conflict detection across text segments, and evaluative monitoring during extended reading.

The thesis also addressed methodological gaps regarding the conceptual distinction between misinformation and disinformation. By testing whether vulnerability to these categories loads onto distinct latent constructs, the findings demonstrated that both types of misleading content are processed through a common vulnerability dimension despite their theoretical distinction based on producer intent. This contribution illustrates the value of empirically testing conceptual classifications against observable cognitive processing patterns rather than assuming that distinctions defined at the level of content production necessarily translate into psychologically separable forms of vulnerability.

A second major methodological objective was the integration of performance-based metacognitive measures into misinformation research. Using working memory and general knowledge tasks with item-level confidence ratings, the thesis operationalized metacognitive bias and sensitivity through objective accuracy-confidence relations rather than self-report measures. This approach allowed overconfidence to be conceptualized as a measurable miscalibration between confidence and performance rather than a subjective belief about competence. Furthermore, the distinction between metacognitive bias and metacognitive sensitivity provided a more fine-grained assessment of monitoring processes than aggregate confidence or accuracy measures alone.

The thesis further extended misinformation methodology by implementing a pre-post feedback design capable of assessing both naturalistic relations and experimentally modifiable metacognitive processes. Although intervention effects on fake news vulnerability

were limited, the design demonstrates how metacognitive processes can be experimentally isolated, manipulated, and assessed longitudinally.

Finally, the thesis addressed the methodological limitations of single-session exposure paradigms through a two-week repeated-exposure design using full-length online articles. This approach allowed examination of cumulative familiarity effects, gradual evaluative drift, and repeated exposure dynamics under conditions more closely approximating digital information ecosystems characterized by algorithmic amplification and recurrent narrative exposure. The inclusion of a mixed-exposure condition additionally provided an initial experimental approximation of informational diversity within algorithmic feeds.

Collectively, these methodological contributions move misinformation research toward greater ecological realism while preserving experimental control and measurement precision.

Practical objectives

Beyond its theoretical and methodological aims, the thesis sought to identify practical implications for interventions targeting health-related fake news vulnerability. A first practical objective concerned whether reducing overconfidence could improve resistance to fake news. Although the metacognitive intervention successfully improved calibration, it did not reduce fake news vulnerability, suggesting that domain-general improvements in confidence calibration may not automatically transfer to complex misinformation evaluation contexts. These findings indicate that metacognitive interventions may need to be directly embedded within news evaluation tasks and explicitly connected to uncertainty during information processing rather than relying on generalized accuracy feedback alone.

A second practical objective involved examining whether informational diversity could buffer the effects of repeated exposure to polarized content. The findings suggested that exposure to mixed-valence informational environments partially attenuated the negative

evaluative shifts produced by repeated exposure to exclusively negative material. At the same time, the results highlighted the disproportionate influence of recency and negativity, suggesting that informational diversity must be sustained and temporally well-distributed to effectively counter cumulative fluency effects. These observations carry important implications for digitally curated informational environments in which engagement-driven algorithms preferentially amplify emotionally negative or conflict-oriented material.

The thesis also demonstrated that explicit and implicit evaluative processes may respond differently to misinformation exposure and intervention efforts. While explicit evaluations proved malleable under repeated exposure, implicit attitudes remained comparatively stable across the studied timeframe. This suggests that many real-world misinformation effects may primarily operate at the level of conscious judgment and expressed belief rather than through rapid restructuring of unconscious associations. Consequently, interventions capable of shifting explicit evaluations may not necessarily produce durable long-term changes in automatic evaluative tendencies.

Taken together, the practical findings suggest that effective interventions against misinformation likely require coordinated approaches simultaneously targeting evaluative monitoring, news media literacy, and diverse structured exposure, to broader informational environments within which credibility judgments can be formed.

References

Allington, D., Duffy, B., Wessely, S., Dhavan, N., & Rubin, J. (2021). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. *Psychological Medicine, 51*(10), 1763–1769. https://doi.org/10.1017/S003329172000224X

Arin, K. P., Koyuncu, M., Spagnolo, N., & Reich, O. F. (2023). The determinants of fake news belief: Evidence from the United Kingdom and Germany. *European Journal of Political Economy, 76*, 102256. https://doi.org/10.1016/j.ejpoleco.2022.102256

Ashley, S., Maksl, A., & Craft, S. (2013). Developing a news media literacy scale. *Journalism & Mass Communication Educator, 68*(1), 7–21. https://doi.org/10.1177/1077695812469802

Bainbridge, T. F., Quinlan, J. A., Mar, R. A., & Smillie, L. D. (2018). Openness/intellect and susceptibility to pseudo-profound bullshit: A replication and extension. *European Journal of Personality, 33*(1), 72–88. https://doi.org/10.1002/per.2186

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition, 3*(1), 2. https://doi.org/10.5334/joc.91

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2021). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science, 17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037/1089-2680.5.4.323

Berezow, A. (2017). A science journalist ranks the 'trustworthiness' of major news outlets. *American Council on Science and Health*.

Bornstein, R. F., & Craver-Lemley, C. (2022). Mere exposure effect. In *Encyclopedia of personality and individual differences*. Springer.

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition, 8*(1), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The Generic Conspiracist Beliefs Scale. *Frontiers in Psychology, 4*, 279. https://doi.org/10.3389/fpsyg.2013.00279

Brown, A. S., & Nix, L. A. (1996). Age-related changes in the tip-of-the-tongue experience. *The American Journal of Psychology, 109*(1), 79–91.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Čavojová, V., Šrol, J., & Jurkovič, M. (2019). Why should we try to think positively about the world? Interrelationships between worldview, cognitive biases, and bullshit receptivity. *Studia Psychologica, 61*(2), 106–121.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*, 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Lawrence Erlbaum.

Cone, J., & Ferguson, M. J. (2015). He did what? The role of diagnosticity in revising implicit evaluations. *Journal of Personality and Social Psychology, 108*(1), 37–57. https://doi.org/10.1037/pspa0000014

De Keersmaecker, J., & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence, 65*, 107–110. https://doi.org/10.1016/j.intell.2017.10.005

De Keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, 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*(2), 204–215. https://doi.org/10.1177/0146167219853844

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(2), 169–187. https://doi.org/10.1080/13546783.2013.854725

DeYoung, C. G., Grazioplene, R. G., & Peterson, J. B. (2012). From madness to genius: The Openness/Intellect trait domain as a paradoxical simplex. *Journal of Research in Personality, 46*(1), 63–78. https://doi.org/10.1016/j.jrp.2011.12.003

Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. O. (2020). A topic-modeling approach to fake news detection. *Information Systems Frontiers, 24*, 1347–1363. https://doi.org/10.1007/s10796-020-10057-8

Ek, S. (2015). Gender differences in health information behaviour: A Finnish population-based survey. *Health Promotion International, 30*(3), 736–745. https://doi.org/10.1093/heapro/dat063

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition. *Perspectives on Psychological Science, 8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Fazio, L. K. (2020). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review, 27*, 185–191. https://doi.org/10.3758/s13423-019-01651-4

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*(5), 993–1002. https://doi.org/10.1037/xge0000098

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review, 26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443. https://doi.org/10.3389/fnhum.2014.00443

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? *ACM SIGMETRICS Performance Evaluation Review, 44*(1), 179–192. https://doi.org/10.1145/2964791.2901462

Galasso, V., Pons, V., Profeta, P., Becher, M., Brouard, S., & Foucault, M. (2020). Gender differences in COVID-19 attitudes and behavior: Panel evidence from eight countries. *Proceedings of the National Academy of Sciences, 117*(44), 27285–27291. https://doi.org/10.1073/pnas.2012520117

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Haim, M., Graefe, A., & Brosius, H.-B. (2018). Burst of the filter bubble? Effects of personalization on the diversity of Google News. *Digital Journalism, 6*(3), 330–343. https://doi.org/10.1080/21670811.2017.1338145

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior, 16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Ireton, C., & Posetti, J. (2018). *Journalism, fake news & disinformation: Handbook for journalism education and training*. UNESCO.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist, 65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Kahan, D. M. (2016). The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. In *Emerging trends in the social and behavioral sciences* (pp. 1–16). Wiley.

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

King, A. J., & Greene, C. M. (2024). Demographic predictors of susceptibility to health misinformation. *Journal of Health Communication*. Advance online publication.

Kurdi, B., & Banaji, M. R. (2017). Reports of the death of the individual difference approach to implicit social cognition may be greatly exaggerated. *Psychological Inquiry, 28*(4), 281–287.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., Tomezsko, D., & Banaji, M. R. (2021). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American Psychologist, 76*(5), 751–768. https://doi.org/10.1037/amp0000764

Lee, J. J., Kang, K.-A., Wang, M. P., Zhao, S. Z., Wong, J. Y. H., O'Connor, S., Yang, S. C., & Shin, S. (2020). Associations between COVID-19 misinformation exposure and belief with COVID-19 knowledge and preventive behaviors: Cross-sectional online study. *Journal of Medical Internet Research, 22*(11), e22205. https://doi.org/10.2196/22205

Little, T. D. (2013). *Longitudinal structural equation modeling*. Guilford Press.

Lindeman, M. (2011). Belief in complementary and alternative medicine: An integrative model of predisposition, perception, and cognition. *Personality and Individual Differences, 51*(5), 575–580. https://doi.org/10.1016/j.paid.2011.05.008

Lobato, E., Mendoza, J., Sims, V., & Chin, M. (2014). Examining the relationship between conspiracy theories, paranormal beliefs, and pseudoscience acceptance among a university population. *Applied Cognitive Psychology, 28*(5), 617–625. https://doi.org/10.1002/acp.3042

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour, 5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Lyons, B. A., Merola, V., & Reifler, J. (2021). Shifting medical misinformation beliefs in the COVID-19 pandemic: The role of social norms and perceived source credibility. *Harvard Kennedy School Misinformation Review*. https://doi.org/10.37016/mr-2020-59

Mækelæ, M. J., Moritz, S., & Pfuhl, G. (2018). Are psychotic experiences related to the perception of profoundness in random statements? *Psychosis, 10*(1), 74–79. https://doi.org/10.1080/17522439.2017.1349827

Miller, T. M., & Geraci, L. (2014). Training metacognition in the classroom: The influence of incentives and feedback on exam predictions. *Metacognition and Learning, 9*, 303–314. https://doi.org/10.1007/s11409-014-9113-7

Mirhoseini, S., Li, J., & Xie, W. (2023). Accuracy nudges reduce misinformation sharing in social media environments. *Computers in Human Behavior, 141*, 107612. https://doi.org/10.1016/j.chb.2022.107612

Naveed, M. A., Shaukat, R., & Mukhtar, S. (2021). COVID-19 misinformation, conspiracy beliefs, and preventive behaviors. *Frontiers in Psychology, 12*, 648845. https://doi.org/10.3389/fpsyg.2021.648845

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme, 17*(2), 141–161. https://doi.org/10.1017/epi.2018.32

Novella, S. (2010). The spectrum of bogus health care. *Science-Based Medicine*.

Nyhan, B., Porter, E., Robertson, R. E., & Jamieson, K. H. (2023). Why do beliefs persist after correction? *Annual Review of Political Science, 26*, 1–20.

O'Rear, A. E., & Radvansky, G. A. (2020). Failure to accept retractions: A contribution to the continued influence effect. *Memory & Cognition, 48*, 127–144. https://doi.org/10.3758/s13421-019-00967-9

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

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*(6), 549–563.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592*(7855), 590–595. https://doi.org/10.1038/s41586-021-03344-2

Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences, 116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

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*(2), 185–200. https://doi.org/10.1111/jopy.12476

Porter, E., Wood, T. J., & Bahador, B. (2018). Can presidential misinformation on climate change be corrected? Evidence from Internet and phone experiments. *Research & Politics, 5*(3). https://doi.org/10.1177/2053168018801374

Rao, T. S. S., & Andrade, C. (2011). The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. *Indian Journal of Psychiatry, 53*(2), 95–96. https://doi.org/10.4103/0019-5545.82529

Robertson, R. E., Mourão, R. R., & Thorson, E. (2023). Misinformation correction and attitude persistence. *Political Communication, 40*(3), 365–384.

Schaewitz, L., Vogt, M., & Bender, B. (2020). Message characteristics and fake news credibility. *Computers in Human Behavior Reports, 2*, 100036.

Scherer, L. D., Pennycook, G., & Rand, D. G. (2020). Who is susceptible to online health misinformation? *American Journal of Public Health, 111*(5), 775–782. https://doi.org/10.2105/AJPH.2020.305889

Spohr, D. (2017). Fake news and ideological polarization: Filter bubbles and selective exposure on social media. *Business Information Review, 34*(3), 150–160. https://doi.org/10.1177/0266382117722446

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news". *Digital Journalism, 6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020). Thinking clearly about causal inferences of politically motivated reasoning. *Nature Human Behaviour, 4*, 396–398. https://doi.org/10.1038/s41562-020-0835-0

Thurstone, L. L. (1947). *Multiple-factor analysis*. University of Chicago Press.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Torreggiani, G. (2025). Cognitive style and political congruence in fake news evaluation. *Journal of Experimental Political Science*. Advance online publication.

Unkelbach, C., & Greifeneder, R. (2013). A general model of fluency effects in judgment and decision making. In *The experience of thinking: How the fluency of mental processes influences cognition and behaviour* (pp. 11–32). Psychology Press.

Van den Bulck, J., & Custers, K. (2010). Television exposure is related to fear of avian flu, an ecological study across 23 member states of the European Union. *European Journal of Public Health, 20*(4), 370–374. https://doi.org/10.1093/eurpub/ckp061

Van Dessel, P., De Houwer, J., Gast, A., & Smith, C. T. (2015). Instruction-based approach-avoidance effects. *Experimental Psychology, 62*(3), 161–169.

Van Dessel, P., Hughes, S., & De Houwer, J. (2019). How do actions influence attitudes? An inferential account of the impact of action performance on stimulus evaluation. *Personality and Social Psychology Review, 23*(3), 267–284. https://doi.org/10.1177/1088868318795730

Vranic, A., Bovan, K., & Jugović, I. (2022). The role of overconfidence in susceptibility to misinformation. *Personality and Individual Differences, 186*, 111353. https://doi.org/10.1016/j.paid.2021.111353

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs, 85*(3), 423–441. https://doi.org/10.1080/03637751.2018.1467564

Wardle, C., & Derakhshan, H. (2017). *Information disorder: Toward an interdisciplinary framework for research and policy making*. Council of Europe.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology Monograph Supplement, 9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848