

When Machines Mislead: Trust, Cognitive Processing, and Belief in AI-Labeled Misinformation

Rajeshwari¹, Hemanth Babu R²

¹Psychology PG Scholar, JAIN (Deemed-to-be University), Bengaluru, India
Email: [pooja.banksy\[at\]gmail.com](mailto:pooja.banksy[at]gmail.com)

²Assistant Professor, JAIN (Deemed-to-be University)
Email: [hemanthbabu.r\[at\]jainuniversity.ac.in](mailto:hemanthbabu.r[at]jainuniversity.ac.in)

Abstract: *Generative artificial intelligence (AI) has increased the speed and scale at which plausible text and other media can be produced, raising questions about how people evaluate information explicitly attributed to AI. This integrative narrative review brings together research on misinformation judgments, online credibility cues, trust in automation, dual-process accounts of reasoning, and Need for Cognition (NFC). The literature suggests that people may use source labels as one cue among several when judging credibility, while trust in AI can shape willingness to rely on AI-generated information. Analytic reasoning is associated with better discrimination between true and false news, but NFC is a motivational disposition toward effortful thinking and should not be treated as a direct measure of analytic ability or heuristic use. Evidence also cautions against assuming that AI disclosure uniformly increases or decreases credibility: outcomes may depend on the message, the source cue, prior expectations, and the task. The review proposes a cautious framework in which AI attribution is interpreted through individual beliefs about AI and evaluated alongside message-level evidence. Direct evidence on the joint roles of trust in AI and NFC in responses to AI-labeled misinformation remains limited. Future research should measure perceived accuracy and truth discernment separately, distinguish source trust from message accuracy, and test these relationships across culturally diverse samples, including India.*

Keywords: AI-generated misinformation; trust in AI; Need for Cognition; source attribution; truth discernment

1. Introduction

Digital platforms expose people to large volumes of information, often with limited time or context for checking accuracy. Misinformation research shows that people can rate false news as plausible and that analytic thinking is associated with better discrimination between true and false headlines [1,2]. Generative artificial intelligence (AI) adds a further dimension: systems can produce fluent, news-like text at scale, while readers may have difficulty distinguishing machine-generated material from human-written content.

Experimental work by Kreps et al. [3] found that participants were often unable to distinguish AI-generated from human-generated news text. This does not mean that AI-generated information is always believed, nor that its source label inevitably changes credibility. Instead, the finding highlights the need to examine how people interpret AI attribution alongside the content and other contextual cues. A label such as “AI-generated” may be interpreted as a sign of technological competence by some readers and as a warning about accuracy or accountability by others.

Research on online credibility, trust in automation, and cognitive processing has developed largely in partly separate areas. Credibility studies examine source and contextual cues; automation research examines trust and reliance; misinformation research examines accuracy judgments and truth discernment. Integrating these strands may clarify why individuals respond differently to the same AI-labeled claim. This review therefore examines how trust in AI, source attribution, and cognitive engagement may relate to belief in

AI-labeled misinformation. It also identifies unresolved questions for future empirical research, rather than treating a proposed relationship between trust and NFC as an established effect.

2. Review Approach

This article uses an integrative narrative review approach to connect findings and concepts across research on online credibility, misinformation, trust in automation, dual-process theories, and AI-generated content. The source set was drawn from the literature and references assembled for the original manuscript. Sources were retained for the present synthesis when they directly informed at least one of the following themes: (a) credibility judgments in online environments; (b) susceptibility to misinformation and truth discernment; (c) trust and reliance on automated systems; (d) cognitive engagement and NFC; or (e) AI-generated content and source disclosure.

The review is conceptual and does not report new participant data, a quantitative meta-analysis, or a systematic database search. The original manuscript did not document a reproducible search strategy, screening flow, or complete eligibility log; consequently, no search counts or claims of exhaustive coverage are reported. Priority is given to identifiable peer-reviewed articles and clearly identified research outputs. A small number of directly relevant preprint findings are described as such. This approach supports conceptual integration but limits claims about completeness and the relative weight of all available evidence.

2.1 Scope and Terminology

Several terms require separation. AI-generated content refers to material produced in whole or in part by an AI system, while AI-labeled content refers to material presented to an audience as AI-generated, regardless of whether the label is accurate. AI-generated misinformation is content that is false or misleading and is produced using AI; AI-labeled misinformation, as used in this review, refers more narrowly to false or misleading content that participants are told was generated by AI. These categories overlap, but they are not identical. A person may react to the label rather than detect the actual origin of the material, and an AI system may produce accurate content as well as inaccurate content.

This distinction matters for causal interpretation. If participants are shown different messages in an AI condition and a human condition, any observed difference may reflect content quality rather than attribution. If the same message is presented with different source labels, the design more directly tests the effect of attribution, although it may introduce demand characteristics if the purpose becomes obvious. Studies should describe precisely what participants see, whether authorship is real or experimentally assigned, and whether the AI system provides an explanation in addition to a label.

The present review focuses on the psychological evaluation of information when AI is salient as a source. It does not assume that all false content generated by AI is intentionally deceptive, nor that every AI-related misinformation event is best explained by individual cognitive bias. Production practices, platform distribution, political and commercial incentives, and the availability of fact-checking also shape the information environment. The individual-level framework developed here addresses one part of a wider socio-technical problem and should be interpreted alongside those structural factors.

3. Theoretical Foundations for Interpreting AI Labels

Dual-process accounts distinguish relatively intuitive judgments from more deliberate, effortful reasoning, although contemporary theories differ in how these processes are defined and related [4]. In online information settings, people cannot investigate every claim in depth. Initial impressions and simple cues may therefore influence judgments, with more extensive evaluation occurring when motivation, ability, time, and relevant evidence permit it. This framework offers a useful way to formulate questions about AI attribution, but it should not be reduced to a rigid claim that all fast judgments are wrong and all deliberative judgments are correct.

An AI label can operate as a cue at the initial stage of evaluation. Its influence may be strengthened when the reader has limited knowledge of the topic or little opportunity to check the claim. However, deliberate processing is not guaranteed to overturn an initial impression. It may confirm it, especially when the reader lacks the knowledge needed to evaluate the underlying evidence. Thus, cognitive engagement and accuracy should

be treated as related but distinct: the amount of thinking a person reports or performs does not by itself establish that the resulting judgment is correct.

The Elaboration Likelihood Model and related dual-process perspectives also suggest that source-related cues and message-based arguments can both contribute to persuasion, with their relative influence depending on involvement and the conditions of processing. In the present context, AI attribution may be processed as a peripheral or heuristic cue when the reader has little time or low involvement. Under other conditions, the label itself may prompt more scrutiny. These are theoretical possibilities, not universal rules, and should be tested in designs that vary the label, message quality, and task conditions independently.

The concept of appropriate reliance adds a useful normative dimension. The goal is not to make people generally more trusting or generally more skeptical, but to help them distinguish between cases in which AI output is useful and cases in which it is unreliable. Schemmer et al. [5] argue that appropriate reliance requires acceptance of correct advice and rejection of incorrect advice. Applied to misinformation, this means that interventions should be evaluated by their effect on accuracy and discrimination, not merely by whether they lower trust in AI.

4. Results of the Integrative Synthesis

4.1 Credibility judgments rely on multiple cues

Credibility is not determined by a message alone. People may draw on source reputation, presentation, familiarity, contextual signals, and their own prior knowledge. Hilligoss and Rieh [6] describe credibility assessment as involving multiple levels, including general beliefs, heuristics, and judgments formed in a particular information context. Metzger and Flanagin [7] similarly explain how cognitive heuristics can help people manage the large volume of information available online. These shortcuts can be useful when time or expertise is limited, but they can also produce inaccurate judgments when cues are weakly related to truth.

AI attribution can be understood as an additional source cue within this broader process. The label may be salient, but its meaning is not fixed. A reader who associates AI with speed, technical knowledge, or consistency may evaluate an AI-labeled message differently from a reader who associates it with errors, opacity, or a lack of accountability. Source disclosure is therefore not equivalent to a direct measure of the message's factual quality. Nor does identifying the author as AI establish whether a claim is true or false.

Research on AI-generated news demonstrates why this distinction matters. Kreps et al. [3] found that AI-generated news text could be difficult for participants to distinguish from human-generated text. The implication is not that AI labels are irrelevant, but that readers may not reliably infer authorship from the text itself. In applied settings, a label may influence judgments differently from actual authorship, and both should be measured when possible.

4.2. Trust in AI is related to reliance, not necessarily accuracy

Trust is an important factor in human interaction with automated systems, but trust and appropriate reliance are not identical. Hoff and Bashir [8] synthesize evidence showing that trust can vary with dispositional, situational, and learned factors. The practical goal is not maximal trust; it is calibrated reliance, in which users accept useful outputs while remaining able to question incorrect ones. A person may have high general trust in AI yet reject a particular false claim, or distrust AI while also rejecting accurate information.

Automation research has documented the risk of automation bias, in which people give undue weight to automated recommendations, as well as algorithm aversion, in which people reduce reliance after observing an algorithm make an error [9,10]. These opposing tendencies show why general attitudes toward AI cannot be mapped directly onto misinformation susceptibility. Trust may shape how a source is initially interpreted, but the final judgment also depends on the claim, available evidence, context, and whether the reader checks the information.

Recent work directly addressing AI-generated misinformation reinforces this caution. Spearing et al. [11] tested source-focused interventions, debunking, and a disclaimer. Their findings indicate that reducing general trust in AI-generated information does not necessarily eliminate the influence of misleading content on reasoning. Trust is therefore one possible contributor to information evaluation, not a sufficient explanation for whether a specific false claim is accepted.

4.3 Analytic reasoning and Need for Cognition are distinct

A substantial misinformation literature links analytic thinking with more accurate truth judgments. Pennycook and Rand [1] reported that susceptibility to partisan fake news was better explained by insufficient reasoning than by motivated reasoning in their studies. Their later review summarizes evidence that analytic thinking and accuracy-focused attention can improve discrimination between true and false information [2]. These findings support the importance of engaging with the accuracy of a claim rather than relying exclusively on familiarity, partisan fit, or source cues.

However, analytic reasoning performance must be distinguished from Need for Cognition. NFC refers to the tendency to engage in and enjoy effortful cognitive activity [12]. It is a motivational disposition, not a direct measure of intelligence, reasoning skill, or actual analytic processing in a particular task. A person high in NFC may be more inclined to examine a claim, but whether this inclination produces accurate judgments depends on knowledge, motivation, task demands, and the evidence available. Similarly, a person low in NFC does not necessarily use heuristics in every situation.

For the present topic, NFC is best treated as a potential individual-difference variable that may influence the extent to which an AI source label is considered or challenged. It should not be assumed that high NFC automatically produces skepticism toward AI or that low NFC necessarily produces belief in AI-generated misinformation. Careful processing may support rejection of false claims when relevant evidence is available, but more thought alone does not guarantee accuracy.

4.4 Misinformation belief and truth discernment are different outcomes

A key measurement issue is the distinction between accepting false information and discriminating between true and false information. Someone who rates both true and false headlines as inaccurate may appear resistant to misinformation if only false-item ratings are examined. However, that response reflects generalized skepticism rather than accurate truth discernment. A stronger design measures perceived accuracy for true and false items separately and examines the difference between them.

Pennycook and Rand [1] used perceived accuracy judgments of real and fake news headlines to examine this distinction. Work on the illusory truth effect also indicates that repeated exposure can increase perceived truth, even when repetition does not provide new evidence [13,14]. These findings show that message-level familiarity and cognitive evaluation can operate alongside source cues. For studies of AI-labeled misinformation, perceived accuracy of false claims should therefore be reported separately from the ability to distinguish true from false claims.

AI-related misinformation may also include persuasive explanations, not merely a label indicating that content was generated by AI. Danry et al. [15], in a preregistered online experiment reported as a preprint, found that deceptive AI-generated explanations could increase belief in false headlines. Their work concerns AI explanations rather than source labels alone, so it provides adjacent rather than direct evidence for the framework proposed here. It nevertheless highlights the importance of separating the effect of authorship attribution from the persuasive content accompanying a claim.

Literature strand	Contribution to the present framework
Online credibility and heuristics	People use source, contextual, and message cues; an AI label is one cue rather than proof of accuracy.
Trust in automation	Trust influences reliance, but appropriate reliance requires accepting correct and rejecting incorrect outputs.
Analytic reasoning and misinformation	Analytic thinking and accuracy attention relate to truth discernment; effects should not be conflated with NFC.
AI-generated content and explanations	AI-generated text can resemble human content; explanations may persuade, but are not equivalent to source labels.

4.5. Source Disclosure Is Not the Same as Message Verification

A recurring problem in public discussions of AI-generated content is the assumption that disclosure resolves the credibility problem. Disclosure provides information about production or authorship, but it does not directly establish the truth value of a claim. A human-authored post can be false, and an AI-generated explanation can contain accurate information. Conversely, an AI-generated message may be misleading even when its style is fluent and its reasoning appears coherent. Source labels should therefore be treated as one input into evaluation rather than as a substitute for evidence.

Disclosure can also have competing effects. It may encourage scrutiny by making readers aware of possible limitations, but it may also create an overly broad inference that AI is either more objective or less trustworthy than humans. If a label is interpreted as a warning, users may reject true content; if it is interpreted as a sign of computational authority, users may accept false content. This possibility underscores the need to measure both acceptance of false information and acceptance of true information. Without this distinction, the direction and practical value of a disclosure effect can be misinterpreted.

Researchers should also distinguish between an AI-generated message, a human message edited with AI, a message attributed to AI, and a message accompanied by an AI-generated explanation. These conditions differ in what the participant is being told and what psychological process may be activated. The source label may alter expectations about authorship, while an explanation may provide persuasive arguments or apparent justification for a claim. Combining all such conditions under the term “AI misinformation” risks obscuring the mechanism under investigation.

4.6 Synthesis of Evidence and the Unresolved Intersection

Across the reviewed literatures, three propositions are comparatively well supported. First, people use multiple cues when judging online credibility, and heuristic reliance can be understandable in information-rich environments [6,7]. Second, trust affects reliance on automation, but the relationship between trust and performance depends on context and calibration [5,8]. Third, analytic thinking, relevant knowledge, and attention to accuracy are associated with better truth discernment, while familiarity and repetition can increase perceived accuracy [1,2,14].

The less established proposition is that trust in AI and NFC interact in a specific way to determine belief in AI-labeled false news. Evidence from each adjacent field creates a rationale for studying this possibility, but it cannot substitute for a direct test. Research on general trust in automation may not predict judgments of news claims; research on analytic reasoning does not directly measure NFC; and studies of AI-generated explanations do not isolate the effect of an AI source label. A careful review should preserve these distinctions rather than combine them into a single claim of demonstrated moderation.

The most defensible conclusion is therefore one of convergence with an important gap. Existing evidence supports the relevance of source cues, trust, and cognitive processing, but the joint pathway linking them to AI-labeled misinformation remains under-specified. A study that tests these variables together could contribute by estimating their separate associations, assessing whether trust predicts perceived accuracy of false claims, and testing whether NFC changes that association. It should also examine truth discernment to determine whether any effect represents improved accuracy or generalized acceptance/rejection.

5. An Integrative Conceptual Framework

The literature supports a cautious framework in which AI attribution, trust in AI, cognitive engagement, and message-level characteristics may jointly contribute to credibility judgments. The proposed pathway is: AI source attribution → interpretation of the source cue → evaluation of the claim and available evidence → perceived accuracy judgment. Trust in AI may shape the interpretation of the source cue, while NFC may influence the motivation to engage in effortful evaluation. Message fluency, familiarity, prior knowledge, and contextual cues may also affect the judgment.

This pathway is conceptual rather than an established causal sequence. The reviewed literature supports individual components—such as the role of source cues in credibility assessment, the relationship between analytic thinking and misinformation discernment, and the importance of trust in automation—but does not establish that NFC moderates the relationship between trust in AI and belief in AI-labeled misinformation. The moderation hypothesis should therefore be tested rather than presented as a conclusion.

The framework also predicts that source trust and message accuracy can diverge. High trust may be beneficial when the information is correct but risky when it encourages uncritical acceptance. Low trust may reduce reliance on false information but can also lead to the rejection of true information. Accordingly, an outcome focused only on willingness to trust AI may miss whether users are actually distinguishing accurate from inaccurate content.

5.1 Testable propositions for future research

The framework yields several propositions that can be evaluated without assuming that the expected effects have already been demonstrated. First, the effect of an AI source label on perceived accuracy should be context-dependent rather than uniformly positive or negative. The direction may vary with prior expectations about AI, message characteristics, and the credibility cues available to the reader.

Second, trust in AI may be associated with greater acceptance of AI-attributed information in some contexts, but this association should not be interpreted as evidence of misinformation susceptibility unless it is shown to apply specifically to false claims. Third, NFC may be associated with a greater willingness to engage in effortful evaluation, but any relationship with accuracy should depend on whether

participants have access to relevant evidence and sufficient knowledge to evaluate it.

Fourth, trust in AI and NFC may interact, but the direction and size of that interaction remain open empirical questions. For example, NFC could reduce reliance on an AI label when the label is inconsistent with available evidence, but it could also have little effect when participants cannot verify the claim. Fifth, the most informative evaluation of interventions should assess truth discernment and appropriate reliance, not merely changes in overall trust or average credibility ratings. These propositions define a research agenda; they are not findings of the present review.

6. Discussion and Implications

For researchers, three design priorities follow. First, studies should distinguish source attribution from actual authorship by varying labels while keeping message content constant where feasible. Second, researchers should measure trust in AI and NFC as distinct constructs and avoid using NFC as a proxy for analytic ability or heuristic processing. Third, outcomes should include both perceived accuracy of false claims and truth discernment across true and false claims. Such designs would make it possible to identify whether an AI label shifts judgments, whether the shift depends on individual differences, and whether any change improves or reduces accuracy.

For educators and platforms, the implication is not simply to encourage trust or distrust in AI. Interventions should support calibrated evaluation: checking claims against independent sources, attending to evidence rather than presentation alone, and recognizing that authorship labels do not verify factual accuracy. Source-focused warnings may help in some circumstances, but they should not substitute for claim-specific evidence and corrections. Interventions should also be assessed for unintended effects, including excessive skepticism toward accurate information.

The main contribution of this integrative review is to connect research areas that explain different parts of information evaluation. Credibility research explains why people use source and contextual cues; trust-in-automation research explains why users vary in their willingness to rely on AI; misinformation research highlights the importance of accuracy judgments and truth discernment; and cognitive-processing research explains why engagement with evidence can matter. No single factor is sufficient to explain responses to AI-labeled misinformation.

6.1. Implications for Measurement and Research Design

A practical design would manipulate source attribution while holding the factual content and wording of each item constant. For example, participants could evaluate the same set of verified true and false headlines under different source-label conditions. If the study instead uses different content in each condition, differences in plausibility, topic familiarity, emotionality, or writing quality may be mistaken for a source-label effect. A balanced set of items and random assignment would help reduce these risks.

The main outcomes should be defined before data collection. Perceived accuracy of false items can index misinformation belief, while truth discernment can be operationalized as the difference between average accuracy ratings for true and false items. These outcomes answer different questions. A source label that lowers ratings for all items may reduce belief in false information without improving discrimination. Conversely, a label could increase confidence in true information while leaving false-item ratings unchanged. Reporting only one aggregate score may conceal such patterns.

Trust in AI should be measured as a construct distinct from trust in a particular AI system, confidence in the message, and perceived expertise of the source. A general trust scale may capture stable attitudes but may not reflect situational trust in the specific message. When feasible, studies can measure general trust before or after the task and separately ask how credible participants found the particular source and claim. The timing of the measures matters: asking about trust before the headline task could prime participants to focus on AI, whereas asking afterward may reduce this priming but could allow the task to influence self-reports.

NFC should likewise be measured and interpreted according to its construct definition. It reflects a preference for effortful thought rather than the actual quality of reasoning in every situation. If researchers want to test analytic ability or cognitive reflection, those constructs should be assessed separately instead of being inferred from NFC. A moderation analysis should be specified in advance, interpreted with confidence intervals, and supported by adequate statistical power. A non-significant interaction should not be treated as proof that cognitive engagement never matters; it may also reflect measurement error, restricted variability, or a context in which the source label has little influence.

6.2. Implications for Public Communication and AI Literacy

AI literacy initiatives should emphasize verification habits rather than a simple rule to trust or distrust machine-generated material. Readers can be taught to identify the factual claim being made, locate independent corroboration, distinguish evidence from confident wording, and check whether the cited source actually supports the claim. This is especially important because fluency and repetition can make information feel familiar or credible without adding evidence. A transparent label can help users understand provenance, but it should be paired with clear accountability and mechanisms for correcting errors.

Platform design also matters. Interfaces can direct attention toward accuracy at the moment users are about to share information, present links to credible sources, or distinguish verified information from unverified claims without implying that automated classification is infallible. Any intervention should be evaluated for both benefits and costs. An intervention that reduces false belief but also reduces acceptance of accurate information may have limited practical value. Measures of truth discernment and calibrated reliance are therefore more informative than changes in generalized trust alone.

For the Indian context, research should account for linguistic and cultural diversity and for variation in access to digital tools. A source label may carry different meanings for people with different levels of familiarity with AI, educational backgrounds, and experiences of online news. Studies should consider whether headlines translated across languages remain equivalent in meaning, plausibility, and emotional tone. They should also avoid treating urban, English-speaking participants as representative of all Indian information users. These considerations make local research important not because the underlying psychological principles must be entirely different, but because the cues and conditions through which they operate may vary.

6.3. Theoretical implications: separating source, message, and person

The framework developed here is useful because it separates three levels of explanation that are often blurred together. The source level concerns who or what is presented as producing the information and how the reader interprets that attribution. The message level concerns features such as plausibility, evidence, familiarity, fluency, and the presence of an explanation. The person level concerns relatively stable orientations and abilities, including trust in AI, NFC, relevant knowledge, and analytic reasoning. These levels may influence one another, but they should not be treated as interchangeable.

For example, if an AI label is associated with higher perceived accuracy, that result alone would not show whether participants trusted AI more, assumed AI to be technically competent, or simply paid more attention to the label. Similarly, if participants with higher NFC show better truth discernment, this would not establish that they were less trusting of AI. They may instead have evaluated evidence more closely, had more topic knowledge, or been less influenced by familiarity. Studies that measure only the source label and a single belief outcome cannot distinguish these possible explanations.

A further distinction is between an effect on average credibility and an effect on discrimination. A label might shift ratings upward for all messages, downward for all messages, or selectively alter ratings for false messages. These patterns have different implications. A general downward shift could indicate increased skepticism without improved accuracy, while selective rejection of false information could indicate better discernment. The framework therefore treats perceived credibility, misinformation belief, and truth discernment as related but non-identical outcomes.

This approach also helps prevent overstatement in interpreting the literature. Studies of human–AI trust often examine advice-taking in decision tasks, whereas misinformation studies typically ask participants to rate claims or headlines. The tasks differ in consequences, participant expertise, and opportunities for feedback. Findings about reliance on an automated decision aid can inform hypotheses about AI-generated news, but they cannot establish that the same effects will occur in social media contexts. A conceptual synthesis should identify such links

as plausible bridges and specify where direct evidence is still needed.

6.4. Research priorities for the next stage of evidence

First, research should test source attribution independently of actual authorship. A factorial design could vary whether a claim is labelled AI-generated or human-written while keeping the same claim content across conditions. Where feasible, a second factor could vary whether the claim is true or false. This would permit researchers to estimate whether an AI label affects perceived accuracy overall and whether it affects true and false claims differently. Participants should be debriefed when deception is necessary, and headline stimuli should be checked for topic familiarity, plausibility, emotionality, and language difficulty.

Second, researchers should examine whether trust in AI predicts judgments beyond general trust in online information and prior attitudes toward technology. Trust can be domain-specific: a person may trust AI for grammar correction but not for health information or news. Measuring general trust alongside task-specific perceived credibility would help establish whether a broad disposition has explanatory value for a particular information judgment. Researchers should also avoid assuming that self-reported trust necessarily predicts actual reliance or accuracy.

Third, studies should examine the conditions under which NFC matters. NFC may predict willingness to engage with a claim, but its relationship with accuracy could depend on the availability of relevant knowledge and the quality of evidence. An interaction between trust and NFC is one reasonable hypothesis, but alternatives should also be considered: NFC may have an independent association with truth discernment, it may be unrelated to judgments under low-involvement conditions, or it may matter only when the task makes verification possible. The theory should therefore generate competing, testable predictions rather than a single assumed direction.

Fourth, future work should investigate mechanisms that occur after a source label is noticed. Researchers could ask participants to report what they thought the label implied, whether they checked the claim, which cues influenced their judgment, and how confident they were. Such measures could help distinguish a source-credibility pathway from a message-evaluation pathway. Self-report measures have their own limitations, especially because people may not accurately describe the processes that influenced them, so they should complement rather than replace behavioral measures.

Fifth, researchers should be cautious about interpreting demographic variables as explanations in themselves. Education, age, and frequency of social media use may relate to familiarity with online information environments, but their effects can differ by context and may be confounded with access, language, and topic knowledge. If these variables are included as controls, their rationale should be specified before analysis. Control variables do not automatically remove bias and should not be added solely because they are available.

Finally, future syntheses would benefit from a transparent search and screening procedure. A subsequent systematic or scoping review could search multidisciplinary databases using combinations of terms for AI-generated content, source disclosure, trust in AI, misinformation, credibility, analytical thinking, and NFC. It should record the search date, databases, inclusion criteria, screening decisions, and reasons for exclusion. A structured evidence table could then distinguish experimental source-label studies from observational work, studies of AI explanations, and broader trust-in-automation research. This would allow future authors to assess how much direct evidence exists for the proposed interaction rather than relying primarily on conceptual proximity.

7. Limitations and Future Research

This review has limitations. It is an integrative narrative synthesis rather than a systematic review or meta-analysis. The source set was drawn from the existing manuscript and was not assembled through a documented, reproducible database search; relevant studies may therefore have been missed. The review also combines evidence from different tasks, including human-versus-AI text comparisons, online credibility judgments, general automation, and AI-generated explanations. These are related but not interchangeable research contexts. Findings from adjacent areas are used to develop the framework, not treated as direct proof of effects specific to AI-labeled misinformation.

Future studies should test the proposed framework directly using experimentally controlled source labels and both true and false information. Trust in AI, NFC, and relevant covariates such as education and social media use can be measured separately. Researchers should preregister hypotheses, use validated measures, report item-level and aggregate outcomes transparently, and test whether effects vary across topic, language, and familiarity with AI. Most importantly, analyses should distinguish a change in belief in false claims from a change in general skepticism or truth discernment.

8. Conclusion

AI-generated content creates new conditions for evaluating information, but its effects cannot be explained by source labels alone. Existing research suggests that online credibility judgments rely on multiple cues, trust influences reliance on automated systems, and analytic reasoning is associated with more accurate misinformation judgments. NFC may be relevant to how people engage with AI-labeled claims, but it is not equivalent to analytic ability and its moderating role remains unresolved. An integrative framework that separates source attribution, trust, cognitive engagement, perceived accuracy, and truth discernment offers a basis for more precise empirical research. The practical goal should be calibrated, evidence-sensitive evaluation rather than either indiscriminate trust in AI or blanket distrust.

References

- [1] Pennycook, G., & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*. 2019;188:39–50. doi:10.1016/j.cognition.2018.06.011.
- [2] Pennycook, G., & Rand, D. G. The psychology of fake news. *Trends in Cognitive Sciences*. 2021;25(5):388–402. doi:10.1016/j.tics.2021.02.007.
- [3] Kreps, S., McCain, R. M., & Brundage, M. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of Experimental Political Science*. 2022;9(1):104–117. doi:10.1017/XPS.2020.37.
- [4] Evans, J. St. B. T., & Stanovich, K. E. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*. 2013;8(3):223–241. doi:10.1177/1745691612460685.
- [5] Schemmer, M., Köhl, N., Benz, C., Bartos, A., & Satzger, G. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 2023:410–422. doi:10.1145/3581641.3584066.
- [6] Hilligoss, B., & Rieh, S. Y. Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management*. 2008;44(4):1467–1484. doi:10.1016/j.ipm.2007.10.001.
- [7] Metzger, M. J., & Flanagin, A. J. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*. 2013;59(Part B):210–220. doi:10.1016/j.pragma.2013.07.012.
- [8] Hoff, K. A., & Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*. 2015;57(3):407–434. doi:10.1177/0018720814547570.
- [9] Dietvorst, B. J., Simmons, J. P., & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*. 2015;144(1):114–126. doi:10.1037/xge0000033.
- [10] Skitka, L. J., Mosier, K. L., & Burdick, M. Does automation bias decision-making? *International Journal of Human-Computer Studies*. 1999;51(5):991–1006. doi:10.1006/ijhc.1999.0252.
- [11] Spearing, E. R., Gile, C. I., Fogwill, A. L., Prike, T., Swire-Thompson, B., Lewandowsky, S., & Ecker, U. K. H. Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science*. 2025;12(6):242148. doi:10.1098/rsos.242148.
- [12] Cacioppo, J. T., Petty, R. E., & Kao, C. F. The efficient assessment of need for cognition. *Journal of Personality Assessment*. 1984;48(3):306–307. doi:10.1207/s15327752jpa4803_13.
- [13] Hassan, A., & Barber, S. J. The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*. 2021;6:38. doi:10.1186/s41235-021-00301-5.

- [14] Pennycook, G., Cannon, T. D., & Rand, D. G. Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*. 2018;147(12):1865–1880. doi:10.1037/xge0000465.
- [15] Danry, V., Pataranutaporn, P., Groh, M., Epstein, Z., & Maes, P. Deceptive AI systems that give explanations are more convincing than honest AI systems and can amplify belief in misinformation. *arXiv*. 2024;2408.00024. <https://arxiv.org/abs/2408.00024>.