AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Bulletin

ISSN: 0033-2909

2026, Vol. 152, No. 3, 349–354
https://doi.org/10.1037/bul0000519

COMMENTARY

Artificial Intelligence as a Partner in Meta-Analysis—Research Agenda, User Recommendations, and Speed–Accuracy Tradeoffs: Commentary on Jansen et al. (2025)

Brooke N. Macnamara
Department of Psychological Sciences, Purdue University

Using a metascience framework for improving meta-analyses, Jansen et al. (2025) tested the accuracy and efficiency of data extraction from primary studies used in meta-analyses with a range of large language models. Efficiency was impressive: Across thousands of studies and hundreds of variables, eight large language models took less than an hour combined to extract hundreds of thousands of data points—work estimated to take a human coder >6,500 hr. Nevertheless, accuracy was inconsistent, ranging from high to low depending on the variable. From these results, Jansen et al. recommended (a) a research agenda for investigating the use of artificial intelligence (AI) for data extraction and (b) methods for using AI as a partner for data extraction when conducting systematic reviews. This commentary expands on recommendations for the research agenda, such as investigating AI-induced bias, the illusion of exploratory depth, and using AI to extract study quality data. This commentary also offers further considerations regarding using AI as a meta-analysis partner, such as how iterative prompts might reduce coding independence. Finally, the commentary discusses speed–accuracy tradeoffs in meta-analyses.

Public Significance Statement
This commentary evaluates the research agenda and use of artificial intelligence (AI) recommendations for meta-analyses by Jansen et al. (2025). This comment adds to the proposed research agenda by recommending investigations of AI-induced bias and AI-extracted study quality data and offers further considerations for using AI as a meta-analysis partner before such processes have been tested and evaluated. This comment contributes to improving meta-analyses in psychological research.

Keywords: meta-analysis, artificial intelligence, systematic reviews, large language models

Meta-analyses summarize findings, test theories, and provide a framework for researchers to develop future investigations. High levels of rigor during data extraction are needed to make valid coding decisions, which, in turn, are necessary to produce reliable and valid meta-analyses (Villiger et al., 2022). Nevertheless, high levels of rigor are time consuming and costly. For example, in one sample, Li et al. (2019) found that extracting variables from primary studies took researchers anywhere from less than 1 hr to greater than 6.5 hr per article. With an aim of reducing the time needed for reliable meta-analytic data extraction, Jansen et al. (2025) investigated the accuracy of multiple large language models (LLMs) compared with the data extracted by human meta-analysts.

Jansen et al.'s (2025) results suggested that using artificial intelligence (AI) to extract data from primary articles compared with human raters reflects a speed–accuracy tradeoff. Using LLMs was substantially faster. However, the accuracy of those extractions varied and was consistently lower than human raters (assuming human raters were accurate). Based on their findings, Jansen et al. recommended (a) a research agenda for investigating the use of AI for data extraction and (b) methods for using AI as a synthesis partner with an aim to improve accuracy while maintaining high efficiency. This commentary offers both an extension to Jansen et al.'s research agenda and further considerations regarding some of the recommended methods that have not yet been tested and evaluated.

Research Agenda for Investigating the Use of AI for Data Extraction

Jansen et al. (2025) proposed a research agenda focused on investigating efficiency and accuracy of AI for extracting data for research syntheses. Specifically, they suggest that researchers examine speed–accuracy tradeoffs of using AI in different roles during data

James E. Pustejovsky served as action editor.
Brooke N. Macnamara https://orcid.org/0000-0003-1056-4996
No data were analyzed. There are no conflicts of interest to disclose. This commentary was not funded, not registered, and no review protocols were prepared.

Brooke N. Macnamara played a lead role in conceptualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to Brooke N. Macnamara, Department of Psychological Sciences, Purdue University, 610 Purdue Mall, West Lafayette, IN 47907, United States. Email: bmacnama@purdue.edu

extraction: as an assistant, an independent second coder, and a judge. This commentary offers extensions to the proposed research agenda.

First, above all else, methodological rigor and accuracy are the highest goals in conducting meta-analyses. To this end, several avenues for research may be fruitful: examining AI-induced bias, AI-induced complacency, and the illusion of exploratory breadth; using AI to examine primary study quality; and research on AI methodologies.

AI-Induced Bias, AI-Induced Complacency, and the Illusion of Exploratory Breadth

Jansen et al. (2025) investigated the potential for systemic bias in using LLMs for data extraction. Finding no differences in data extraction accuracy depending on sociocultural context, they conclude that using AI for information retrieval, rather than generation, is unlikely to suffer from bias embedded in the AI tool from training data. However, there are other forms of bias, complacency, and illusions that may be productive areas of investigation.

AI-Induced Bias

Automation bias occurs when users favor information coming from an automated system over a human expert, even when it is less accurate (Manzey et al., 2012; Mosier et al., 1997). Automation bias can influence how users assess a situation, search for patterns, or the frequency with which they cross-check information (Manzey et al., 2012; Rezazade Mehrizi et al., 2023; Smith et al., 1997).

Notably, the types of systems studied when initially identifying automation bias are rudimentary compared with AI. Non-AI automated systems are governed by nonadapting rules where the same input produces the same output from the system in each instance. In contrast, AI systems dynamically assess and conform to new inputs, with many designed to mimic expert knowledge and decision making. For these reasons, people may be especially susceptible to AI-induced bias—overconfidently favoring information coming from an AI over a human, even when it is less accurate (Horowitz & Kahn, 2024; Jones-Jang & Park, 2023).

Just as automation bias can influence how users assess a situation, search for patterns, or the frequency with which they cross-check information, AI-induced bias may influence the user's attention and decision making without the user's awareness (Macnamara et al., 2024; Wang et al., 2023). For example, even using AI to locate key information in an article for the human to extract may bias the researcher to prioritize the information highlighted by an AI over unhighlighted information. This bias may be compounded when asking an LLM to go one step further and extract the key information for a researcher to evaluate, further removing the researcher from the source material. Of note, the more beneficial that users perceive the AI system to be, the more likely they are to agree with inaccurate information presented by the AI (Küking et al., 2024).

To extend Jansen et al.'s (2025) proposed research agenda, one recommendation is to examine how AI-induced bias affects meta-analytic decisions differently depending on the AI's role as an assistant, an independent second coder, or a judge. By definition, an assistant helps a more senior person, typically with greater decision-making power and expertise (here, the meta-analyst). The more domain expertise an individual has, the less they are likely to rely on algorithms and decision aids (Parkes, 2017). As such, the meta-analyst may be less likely to defer to an AI assistant over a higher ranking human independent coder or their own judgment.

In contrast to an assistant, the other two proposed roles—AI as an independent coder or as a judge—imply greater authority and responsibility. As such, meta-analysts may be more prone to deferring to the AI's decisions, reducing the time spent on their own search process, or limiting cross-checking. Jansen et al. (2025) provided clear and strong recommendations for meta-analysts to take responsibility for all final decisions in their work, regardless of the AI's role. Though an agreeable aim, counterintuitively, some research has shown that the more people feel responsible for the outcome of a task, the more they rely on an algorithm's decisions over their own (Van Dongen & Van Maanen, 2013).

The hypothesis that deference may depend on the AI's role could be tested by randomly assigning multiple skilled human researchers to use LLMs in one of four capacities: AI as an assistant, AI as a judge, AI as an independent second coder, and no AI. The researchers could work independently on a new systematic review or seek to replicate the coding of an existing systematic review that was (presumably) originally coded by humans. Researchers working with AI would record time durations, their prompts for the AI, the AI output, the steps in their own search, and the source of each data point.

If AI-induced bias is an issue, then researchers working with an AI, perhaps particularly as a judge or an independent coder, will have more inaccuracies in their decisions compared with the no-AI group. To help ensure that the no-AI group is the "gold standard" for accuracy, the extracted information could be compared with the original meta-analysis if conducting a replication or submitted to an additional evaluation where independent researchers attempt to uncover errors and verify decisions of the original no-AI group. Only by empirical investigations can we determine (a) if AI-induced bias exists in this process, and for which roles, (b) other influential factors, and (c) best practices to mitigate AI-induced bias.

AI Aversion

On the other side of automation bias is algorithmic aversion. Algorithmic aversion occurs when people avoid relying on algorithms even when algorithms outperform humans (Dietvorst et al., 2015; Promberger & Baron, 2006). Algorithmic aversion occurs for multiple reasons, such as that humans, unlike most algorithms, can explain their decision process, aim to be accurate to protect their reputation, and are viewed as capable of taking responsibility (Castelo et al., 2019; Önkal et al., 2009; cf. algorithmic appreciation: when people prefer algorithmic outputs over humans; Logg et al., 2019). Trust in algorithms depends on the algorithms' performance, beliefs about the algorithm's performance, and general comfort with using an algorithm for a task typically and historically carried out by humans (Castelo et al., 2019). Like automation bias, algorithmic aversion can be extended to AI—when people distrust, reject, or avoid using AI even when the AI outperforms humans.

A goal is to find a balance between AI bias and AI aversion: where humans rely on AI when it meets or exceeds the accuracy of humans and refrain from using AI when it is less accurate than humans, at least for tasks where accuracy is paramount, such as meta-analyses. When the goal is to produce valid and reliable syntheses, accuracy needs to be prioritized over efficiency. As methods are developed to improve data extraction accuracy, such as those suggested by Jansen et al. (2025), LLM data extraction may meet or exceed the accuracy

of human coders in the future. In this case, a research agenda for reducing AI aversion might prove beneficial.

One approach may be to ask AI to provide a detailed explanation for decisions and portions of the article searched, as the decision process is often obscured, leading to distrust (Cadario et al., 2021; Morewedge, 2022). That said, while more detailed explanations promote trust, they can also lead to overreliance: Less detailed explanations can prompt people to question a system's reliability (Bussone et al., 2015). Future researchers may seek to determine the optimal level of detail in an explanation to promote use without overreliance.

Another approach may be to develop ways to disambiguate evaluation criteria (Morewedge, 2022). Clear, objective evaluation criteria reduce algorithm aversion (Cadario et al., 2021; Castelo et al., 2019; Morewedge, 2022). Furthermore, clear, objective evaluation criteria—ideally determined a priori and preregistered—have the additional benefit of potentially increasing accuracy, regardless of whether raters are human, AI, or a combination.

AI-Induced Complacency

Related to automation bias, automation complacency occurs when users rely on an automated system while failing to adequately evaluate the system's performance (Risko & Gilbert, 2016). Automation compliance can lead to the introduction of undetected errors and anomalous conditions. Though relying on a trusted system has benefits, research suggests that people may learn to become careless (Saadeh et al., 2025), relying on automated systems even when doing so increases errors (Parasuraman & Manzey, 2010; Wickens & Dixon, 2007). Like with AI-induced bias, people may be especially susceptible to AI-induced complacency—relying on AI while failing to adequately evaluate the system's performance (Saadeh et al., 2025). Further, reliance on AI may hamper one's skills even when not using the AI, further reducing the user's ability to evaluate a system and detect its errors (Macnamara et al., 2024).

Jansen et al. (2025) took care to encourage meta-analysts to take responsibility for the AI's output. A direction for research may be to develop methods for engaging in meaningful human control. For example, future research directions may test ways for the user to evaluate a system and detect its errors and omissions. As another example, future research directions may include investigating how meta-analysts can optimize iterative prompts to increase accuracy and limit bias. Likewise, another direction is to develop guidelines for transparency and reproducibility in this process. Though initial recommendations are beginning to take shape (see Thomas et al., 2025), methods for how best to take responsibility for the AI's outputs have yet to be evaluated and are an excellent avenue for investigation.

Illusion of Exploratory Breadth

The rise of AI use in science has enormous potential for accelerated productivity; nevertheless, accelerated processes carry a risk of decreasing scientists' understanding of those processes and the ultimate outcomes. The illusion of exploratory breadth is a type of metacognitive error stemming from an incorrect belief that one has explored the full problem space, when they have actually explored a narrower space presented to them by the AI (Messeri & Crockett, 2024). In particular, using AI for data extraction may make meta-analysts vulnerable to believing they are considering all possible information rather than only what is available through the LLM.

A careful meta-analyst who is searching for information in a primary study will read not only the primary study article but also the supplementary materials to ensure they are aware of the entire knowledge set associated with the primary study. Further, they may follow links in the article to repositories that contain preregistrations, data, or study materials for further information. They will cross check with other versions of the primary study text, such as preprints or if the study was originally reported in a dissertation that may contain more details. In contrast, an LLM extracting from only the primary study pdf can give the illusion that it has considered the whole knowledge set for that study. An extension to Jansen et al.'s (2025) proposed research agenda may be to measure this illusion and its impact and then to develop methods for evaluating auxiliary study content in conjunction with AI extraction.

Assessing Risk of Bias and Evaluating Primary Study Quality

Another potential extension to Jansen et al.'s (2025) proposed research agenda is to investigate how AI might be able to assess primary study quality. Though Jansen et al. utilized risk of bias criteria in the meta-analyses, this information was used to help extract the timeframe, outcomes, population, interventions, context, study design, and moderators from the primary studies—evaluating risk of bias was outside the scope of the investigation.

Researchers have recently begun testing LLMs for evaluating risk of bias. In a study of randomized clinical trials, Lai et al. (2024) found that two LLMs' accuracy of risk of bias had a wide range depending on the domain. ChatGPT ranged from 56.7% accuracy for assessing random sequence generation to 96.7% accuracy for assessing blinding clinicians to condition. Claude ranged from 80.00% for assessing random sequence generation to 98.3% accuracy for assessing blinding patients to condition. Fifty-seven percent of the evaluative errors were due to an erroneous judgment despite correctly identifying the rationale; the remaining 43% of errors were due to the LLM misidentifying or failing to identify the critical evidence in the original study.

In an analysis of cohort studies, Xia et al. (2025) also found that three LLMs' (Moonshot-v1-128k, ChatGPT-4o, and DeepSeek-V3) accuracy of risk of bias had a considerable range depending on the domain. The models' highest accuracies were 91.67%–93.33% for assessing outcome confidence, outcome baseline, and the completeness of cohort follow-up. Accuracy judgments for assessing cointervention similarity, which is the ability to isolate the critical variable of interest, was the poorest across all three models: 66.67%, 68.33%, and 70.00%, respectively.

Related to assessing risk of bias is evaluating the quality of the primary studies. Evaluating the quality of primary studies is a critical step necessary for establishing the validity of the inferences drawn from the meta-analysis (Scherer & Emslander, 2025; Tod et al., 2022; Valentine, 2019; Valentine & Konstantopoulos, 2016). Yet, this step is often neglected in meta-analyses from psychology.

One reason for this neglect may be due to how time consuming the process of evaluating primary study quality can be. Another reason may be due to the lack of guidance on how to identify and assess primary study quality (Scherer & Emslander, 2025). Such guidance is particularly difficult to generalize due to different standards in study design, implementation, reporting, and aims across disciplines and contexts (Hohn et al., 2019; Scherer & Emslander, 2025).

If meta-analysts prioritize efficiency, then the additional time dedicated to assessing primary study quality, as well as development of guidance for assessing primary study quality, may be further disincentivized.

A program of research that incorporates AI could be developed to establish, evaluate, and implement guidelines for assessing primary study quality. As part of these investigations, researchers may need to consider the following: First, how are the criteria justified and what type of information is necessary to determine if a study has met a particular quality criterion? Second, how will the information be evaluated, ensuring valid and reliable decision making? Third, how can the criteria and decision-making process be sufficiently transparent? A lack of rigor in any stage or type of extraction can lead to erroneous meta-analytic results or signal acceptance of low standards in meta-analytic research (Villiger et al., 2022). Any improvements in primary study quality evaluation in meta-analyses advance metascience.

That said, LLMs' assessment capabilities largely depend on the clarity of the original text. They currently lack the ability to fully consider context, draw on additional resources or professional knowledge, or reason about unclear statements, leading to decision errors (Tang et al., 2023; Xia et al., 2025). Descriptions in the original text used to evaluate risk of bias and original study quality may be more variable than descriptions of timeframes, outcomes, populations, interventions, contexts, study designs, and moderators. Any program of research for improving evaluations of risk of bias and original study quality will need to consider potential limitations of both human experts and LLMs.

Assessing User Recommendations for Extracting Data With AI

Jansen et al. (2025) offered multiple recommendations for using AI to extract data for meta-analysts. Several were based on their findings. For example, they recommend using ensembles of LLMs over single models, which, based on their results, demonstrated higher accuracy. Metascience research may benefit from replications and tests with various models to confirm these results.

Other recommendations by Jansen et al. (2025) have yet to be empirically tested. These recommendations may lead to improved accuracy of extracted effect sizes and variables. Testing the soundness of these recommendations offers a worthwhile direction for implementing a research program on AI for data extraction. Next, the untested recommendations are described, along with additional considerations regarding implementing these practices before their methods have been empirically supported.

Iterative Optimization

Though iterative optimization will almost certainly improve accuracy (Jansen et al., 2025), iterative optimization could lead to bias for some types of information extraction. By iteratively adjusting prompts based on the LLM's output, the user may be training the AI to agree with them, as LLMs can have tendencies to prioritize user approval over objective truth (AI sycophancy; Turner & Eisikovits, 2026).

Iterative prompts may also stray from recommendations of having two independent coders (Cooper, 2017; Lakens et al., 2016; Orwin & Vevea, 2009). Though some recommendations involve the coders engaging in periodic discussions to reach consensus (Orwin & Vevea, 2009), it is unclear if such discussions with an LLM would produce the same results as it would among human coders. For example, given potential AI sycophancy (Turner & Eisikovits, 2026) and AI social desirability (Salecha et al., 2024), it may be the case that when there is disagreement, the LLM will seek to conform over disagreeing with an idiosyncratic interpretation or pointing out a detail the user has missed. If the AI is serving in the role as a judge, iterative prompts to shape decisions may reduce the validity of their role as an objective arbiter.

In addition, the user may shift their own perceptions and decisions through an iterative process. The iterative process of reading outputs and redescribing features for extraction could lead to unrealized changes from initial definitions. Researchers implementing this recommendation may wish to identify and preregister specific definable study characteristics and variables prior to beginning data extraction, taking care to reference these definitions throughout the iterative optimization process.

Using an Additional Coder Only When the AI and Human Disagree

One approach offered for maximizing efficiency while implementing human verification is to bring in a second human coder only when the primary coder disagrees with the majority of LLMs on a data point (Jansen et al., 2025). In addition to this approach not yet being tested, two further considerations are offered.

First, there are multiple decision points for how to work with an AI that might alter the effectiveness of this approach. One example decision point is determining who is engaging with the LLMs. If the primary human coder is also the person iteratively optimizing prompts in the LLMs, the primary coder might be reinforcing agreement with their coding rather than independently verifying it. In this case, coding will primarily reflect the primary coder's perspective rather than independent agreement between a human and AI system. If only bringing in a second coder when the human and AI disagree, an additional human coder's perspective will rarely be considered. An empirical question is how decisions differ depending on the role of the person optimizing the prompts, as well as the role of the LLMs: independent coder, judge, or assistant.

A second concern is related to the experience of the second human coder who only steps in as needed. Coding is a technically demanding step in a meta-analysis, where meta-analysts face numerous decision points that affect model outcomes (Hunt, 1997; Villiger et al., 2022). If a second coder is only asked to occasionally engage in extraction (when the coder and AI disagree), the second human coder will have less experience with the body of literature. Coders need ample experience with the literature to properly understand study content and how it should be coded (Cooper, 2017; Lipsey & Wilson, 2001; Villiger et al., 2022). A quirk found in one primary study (e.g., different descriptions in different parts of the article, degrees of freedom that do not match reported sample sizes) can lead to further investigation for clarity and reinforce checking for similar quirks in other articles. A coder who is fully engaged in the extraction process may be more likely to develop a rigorous process to find, identify, and verify data points than one who only engages when a coder and AI disagree.

Conclusion

This commentary describes potential extensions to Jansen et al.'s (2025) recommended research agenda for using AI in meta-analyses. These suggestions include investigating potential AI-induced bias, AI-induced complacency, and the illusion of exploratory breadth. Another suggestion is to investigate methods for assessing primary study quality. Further, meta-analysts are encouraged to take part in the research agenda put forth by Jansen et al., such as comparing time savings and error rates of AI across different roles: as an assistant, an independent coder, and a judge. Jansen et al. should be applauded for promoting a research agenda for meta-analyses; a goal of this commentary is to offer further avenues for such a program of research.

This commentary also offers additional considerations regarding using AI as a partner in meta-analyses. For example, researchers may need to consider how to iteratively optimize prompts without biasing the AI and without biasing the meta-analyst. Likewise, researchers may need to examine how a full-time human second coder compares with a second human who only codes during AI-primary human coder disagreements. AI may appear to promise improvements in productivity and objectivity, but until methods have been validated to meet (or exceed) the accuracy of current methods, great care is warranted.

Meta-analyses are time consuming and involve large amounts of work, but they also have impact: Meta-analyses can test theories, guide research directions, and influence policy. For this reason, meta-analysts' highest responsibility is to produce accurate results. To this end, when choosing methodological approaches for data extraction that offer a speed–accuracy tradeoff, it is imperative that rigor triumph over efficiency.

References

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *International Conference on Healthcare Informatics* (pp. 160–169). IEEE. https://doi.org/10.1109/ICHI.2015.26

Cadario, R., Longoni, C., & Morewedge, C. K. (2021). Understanding, explaining, and utilizing medical artificial intelligence. *Nature Human Behaviour*, *5*(12), 1636–1642. https://doi.org/10.1038/s41562-021-01146-0

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Cooper, H. (2017). *Research synthesis and meta-analysis: A step-by-step approach* (5th ed.). SAGE Publications. https://doi.org/10.4135/9781071878644

Dietvorst, B., Simmons, J. P. & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Hohn, R. E., Slaney, K. L., & Tafreshi, D. (2019). Primary study quality in psychological meta-analyses: An empirical assessment of recent practice. *Frontiers in Psychology*, *9*, Article 2667. https://doi.org/10.3389/fpsyg.2018.02667

Horowitz, M. C., & Kahn, L. (2024). Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. *International Studies Quarterly*, *68*(2), Article sqae020. https://doi.org/10.1093/isq/sqae020

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. Sage Publications. https://www.jstor.org/stable/10.7758/9781610442961

Jansen, T., Liebenow, L. W., Mertens, U., Schmidt, F. T. C., Lohmann, J. F., Fleckenstein, J., & Meyer, J. (2025). Data extraction by generative artificial intelligence: Assessing determinants of accuracy using human-extracted data from systematic review databases. *Psychological Bulletin*, *151*(10), 1280–1306. https://doi.org/10.1037/bul0000501

Jones-Jang, S. M., & Park, Y. J. (2023). How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. *Journal of Computer-Mediated Communication*, *28*(1), Article zmac029. https://doi.org/10.1093/jcmc/zmac029

Küking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J. O., Moelleken, M., Erfurt-Berge, C., Kutza, J-O., Babitsch, B., & Busch, D. (2024). Automation bias in AI-decision support: Results from an empirical study. In R. Röhrig, N. Grabe, U. H. Hübner, K. Jung, U. Sax, C. O. Schmidt, M. Sedlmayr, & A. Zapf (Eds.), *German medical data sciences 2024* (pp. 298–304). IOS Press. https://doi.org/10.3233/SHTI240871

Lai, H., Ge, L., Sun, M., Pan, B., Huang, J., Hou, L., Yang, Q., Liu, J., Liu, J., Ye, Z., Xia, D., Zhao, W., Wang, X., Liu, M., Ronjan Talukdar, J., Jinhui Tian, J., Yang, K., & Estill, J. (2024). Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Network Open*, *7*(5), Article e2412687. https://doi.org/10.1001/jamanetworkopen.2024.12687

Lakens, D., Hilgard, J., & Staakes, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC Psychology*, *4*(1), Article 24. https://doi.org/10.1186/s40359-016-0126-3

Li, T., Saldanha, I. J., Jap, J., Smith, B. T., Canner, J., Hutfless, S. M., Branch, V., Carini, S., Chan, W., De Bruijn, B., Wallace, B. C., Walsh, S. A., Whamond, E. J., Murad, M. H., Sim, I., Berlin, J. A., Lau, J., Dickersin, K., & Schmid, C. H. (2019). A randomized trial provided new evidence on the accuracy and efficiency of traditional vs. electronically annotated abstraction approaches in systematic reviews. *Journal of Clinical Epidemiology*, *115*, 77–89. https://doi.org/10.1016/j.jclinepi.2019.07.005

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Macnamara, B. N., Berber, I., Çavuşoğlu, M. C., Krupinski, E., Nallapareddy, N., Nelson, N. E., Smith, P. J., Wilson-Delfosse, A. L., & Ray, S. (2024). Does using artificial intelligence assistance accelerate skill decay and hinder skill development without performers' awareness? *Cognitive Research: Principles and Implications*, *9*, Article 46. https://doi.org/10.1186/s41235-024-00572-8

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, *627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Morewedge, C. K., (2022). Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, *26*(10), 824–826. https://doi.org/10.1016/j.tics.2022.07.007

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1997). Automation bias: Decision making and performance in high-tech cockpits. *International Journal of Aviation Psychology*, *8*(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, *22*(4), 390–409. https://doi.org/10.1002/bdm.637

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 177–203). Sage Publications.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

Parkes, A. (2017). The effect of individual and task characteristics on decision aid reliance. *Behaviour & Information Technology*, *36*(2), 165–177. https://doi.org/10.1080/0144929X.2016.1209242

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468. https://doi.org/10.1002/bdm.542

Rezazade Mehrizi, M. H., Mol, F., Peter, M., Ranschaert, E., Dos Santos, D. P., Shahidi, R., Fatehi, M., & Dratsch, T. (2023). The impact of AI suggestions on radiologists' decisions: A pilot study of explainability and attitudinal priming interventions in mammography examination. *Scientific Reports*, *13*(1), Article 9230. https://doi.org/10.1038/s41598-023-36435-3

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Saadeh, M. I., Janhonen, J., Beer, E., Castelyn, C., & Hoffman, D. N. (2025). Automation complacency: Risks of abdicating medical decision making. *AI and Ethics*, *5*, 5783–5793. https://doi.org/10.1007/s43681-025-00825-2

Salecha, A., Ireland, M. E., Subrahmanya, S., Sedoc, J., Ungar, L. H., & Eichstaedt, J. C. (2024). *Large language models show human-like social desirability biases in survey responses*. arXiv. https://doi.org/10.48550/arXiv.2405.06058

Scherer, R., & Emslander, V. (2025). Utilizing primary study quality in meta-analyses in psychology: A step-by-step tutorial. *Psychological Methods*. Advanced online publication. https://doi.org/10.1037/met0000751

Smith, P. J., McCoy, E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man and Cybernetics*, *27*(3), 360–371. https://doi.org/10.1109/3468.568744

Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., Xu, Z., Ding, Y., Durrett, G., Rousseau, J. F., Weng, C., & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, *6*, Article 158. https://doi.org/10.1038/s41746-023-00896-7

Thomas, J., Flemyng, E., Noel-Storr, A., Moy, W., Marshall, I. J., Hajji, R., Jordan, Z., Aromataris, E., Mheissen, S., Clark, A. J., Jemioło, P., Saran, A., Walker, V. R., Angrish, M., Macura, B., Spijker, R., Haddaway, N., Kusa, W., Chi, Y., … Goretzko, J. (2025). *Responsible use of AI in Evidence Synthesis (RAISE)1: Recommendations for practice* [Unpublished project]. https://osf.io/fwaud/files/mrzuj

Tod, D., Booth, A., & Smith, B. (2022). Critical appraisal. *International Review of Sport and Exercise Psychology*, *15*(1), 52–72. https://doi.org/10.1080/1750984X.2021.1952471

Turner, C., & Eisikovits, N. (2026). *Programmed to please: The moral and epistemic harms of AI sycophancy*. SSRN. https://doi.org/10.2139/ssrn.6117867

Valentine, J. C. (2019). Incorporating judgments about study quality into research syntheses. In H. Cooper, L. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 129–140). Russell Sage Foundation. https://doi.org/10.7758/9781610448864

Valentine, J. C., & Konstantopoulos, S. (2016). *Using systematic reviews and meta-analyses to inform public policy decisions*. Commissioned article prepared for the Committee on the Use of Economic Evidence to Inform Investments in Children, Youth, and Families: The National Academies of Sciences, Engineering and Medicine. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_171853.pdf

Van Dongen, K., & Van Maanen, P. P. (2013). A framework for explaining reliance on decision aids. *International Journal of Human–Computer Studies*, *71*(4), 410–424. https://doi.org/10.1016/j.ijhcs.2012.10.018

Villiger, J., Schweiger, S. A., & Baldauf, A. (2022). Making the invisible visible: Guidelines for the coding process in meta-analyses. *Organizational Research Methods*, *25*(4), 716–740. https://doi.org/10.1177/10944281211046312

Wang, D. Y., Ding, J., Sun, A. L., Liu, S. G., Jiang, D., Li, N., & Yu, J. K. (2023). Artificial intelligence suppression as a strategy to mitigate artificial intelligence automation bias. *Journal of the American Medical Informatics Association*, *30*(10), 1684–1692. https://doi.org/10.1093/jamia/ocad118

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Xia, D., Lai, H., Zhao, W., Huang, J., Liu, J., Ye, Z., Liu, J., Sun, M., Hou, L., Pan, B., & Ge, L. (2025). Assessing risk of bias of cohort studies with large language models. *Research Synthesis Methods*, *16*(6), 990–1004. https://doi.org/10.1017/rsm.2025.10028

Received December 1, 2025
Revision received February 13, 2026
Accepted March 18, 2026 ■