AI & SOCIETY
https://doi.org/10.1007/s00146-025-02422-7

OPEN FORUM

Exploring automation bias in human–AI collaboration: a review and implications for explainable AI

Giuseppe Romeo[1] · **Daniela Conti**[1]

Received: 13 March 2025 / Accepted: 2 June 2025

Abstract
As Artificial Intelligence (AI) becomes increasingly embedded in high-stakes domains such as healthcare, law, and public administration, automation bias (AB)—the tendency to over-rely on automated recommendations—has emerged as a critical challenge in human–AI collaboration. While previous reviews have examined AB in traditional computer-assisted decision-making, research on its implications in modern AI-driven work environments remains limited. To address this gap, this research systematically investigates how AB manifests in these settings and the cognitive mechanisms that influence it. Following PRISMA 2020 guidelines, we reviewed 35 peer-reviewed studies from SCOPUS, ScienceDirect, PubMed, and Google Scholar. The included literature, published between January 2015 and April 2025, spans fields such as cognitive psychology, human factors engineering, human–computer interaction, and neuroscience, providing an interdisciplinary foundation for our analysis. Traditional perspectives attribute AB to over-trust in automation or attentional constraints, resulting in users perceiving AI-generated outputs as reliable. However, our review presents a more nuanced view. While confirming some prior findings, it also sheds light on additional interacting factors such as, AI literacy, level of professional expertise, cognitive profile, developmental trust dynamics, task verification demands, and explanation complexity. Notably, although Explainable AI (XAI) and transparency mechanisms are designed to mitigate AB, overly technical, cognitively demanding, or even simplistic explanations may inadvertently reinforce misplaced trust, especially among less experienced professionals with low AI literacy. Taken together, these findings suggest that although explanations may increase perceived system acceptability, they are often insufficient to improve decision accuracy or mitigate AB. Instead, user engagement emerges as the most feasible and impactful point of intervention. As increased verification effort has been shown to reduce complacency toward AI mis-recommendations, we propose explanation design strategies that actively promote critical engagement and independent verification. These conclusions offer both theoretical and practical contributions to bias-aware AI development, underscoring that explanation usability is best supported by features such as understandability and adaptiveness.

Keywords Automation bias · Human–AI collaboration · Trust calibration · Explainable AI (XAI) · Hybrid intelligence · Anchoring effect

1 Introduction

The tendency to over-rely on automated systems has its roots in a cognitive phenomenon known as automation bias (AB). As Artificial Intelligence (AI) continues to influence critical fields such as healthcare, law, and public administration, the need to unravel the mechanisms of AB has grown increasingly urgent (Romeo and Conti 2024). To the best of our knowledge, existing reviews have focused on AB in relation to traditional computer-based aids rather than AI-driven systems (see Goddard et al. 2012; Lyell and Coiera 2017). This paper aims to assess AB specifically within the context of human–AI collaboration. To achieve this, we integrate findings from the studies identified through the PRISMA method into human–AI decision-making research. This approach helped us develop deeper research questions. It also provides a comprehensive and up-to-date perspective on AB in contemporary AI-assisted environments.

✉ Daniela Conti
daniela.conti@unict.it

Giuseppe Romeo
uni390947@studium.unict.it

[1] Department of Humanities, University of Catania, Catania, Italy

Published online: 03 July 2025

1.1 What is human–AI decision-making?

The advance of AI technologies has led to their increasing integration across various job sectors to support decision-making processes. In high-stakes domains such as justice, business, and healthcare, fully automated solutions are often neither practical nor desirable due to safety, ethical, and legal concerns. Conversely, relying solely on manual methods may result in inaccuracies and wasted time. To address these challenges, AI assistance has been proposed to augment human decision-making by offering predictions or recommendations for each decision task. This scenario is commonly referred to as *Human–AI Decision-Making*, alongside related terms such as *Human–AI Collaboration* and *Human–AI Teaming* (Lai et al. 2023).

This vision of accomplishing complex goals by combining human and AI to collectively achieve superior results has also been described as Hybrid Intelligence (HI) systems (Dellermann et al. 2021). This interaction paradigm assumes that collaboration between humans and AI can outperform their independent abilities thanks to the complementarity of their skill sets: AI surpasses human performance levels in specific tasks such as data analysis and image recognition (e.g., cancer detection), whereas it still struggles with complex decision-making, adaptability in dynamic environments, and tasks requiring common sense, areas where humans excel (Dellermann et al. 2021).

1.1.1 Trust calibration and other challenges in hybrid decision-making

Trends in the literature on human–AI interactions vary in their scope of decision-making tasks, but predominantly focus on one aspect: trust (Lai et al. 2023). Trust is arguably the most critical variable in human–automation interaction (Wickens et al. 2021). In the automation context, trust is defined as «the attitude that an (automated) agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability» (Lee and See 2004: p.51). To prevent both automation misuse, where individuals over-rely on systems beyond their capability, and automation disuse, that is the underutilization of technology due to distrust, trust calibration is essential in the design of automated systems. This involves aligning the level of trust with the system's actual capabilities (i.e., system's trustworthiness) to avoid blind reliance on AI, especially since AI systems are not immune to errors (Lee and See 2004). In essence, trust reflects an affective and cognitive evaluation of beliefs about system competence. Proper trust calibration, or appropriate reliance, occurs when users choose to rely on automation when it provides accurate information, and avoid reliance when it does not.

Several factors shape trust, with system reliability being the most influential. Highly reliable systems tend to foster trust and reliance, while failures (particularly the first failure) can significantly erode it. System complexity also undermines trust, as users struggle to understand processes (the "black box effect"). Other key influences include system transparency and individual differences: user's experience, confidence, and familiarity with automation play vital roles in determining trust levels (Wickens et al. 2021).

However, miscalibrated trust can result in overreliance, or over-trust. This occurs when users assume that automation will operate flawlessly, failing to critically evaluate its performance, that is when user's trust overcomes the system's actual accuracy and trustworthiness. In situations with high workload, users may reallocate attention to manual tasks, neglecting the automation. Such behavior increases the risk of delayed detection of automation errors. A direct consequence of this behavioral pattern is the cognitive error of AB (Wickens et al. 2021). Therefore, the effective realization of the potential benefits of human–AI collaboration entails addressing three primary challenges (Steyvers and Kumar 2024):

Understanding and improving human–AI complementarity to develop AI systems that can complement human decision-making rather than simply replace it. This entails identifying areas where AI and human capabilities overlap or differ and leveraging those to achieve superior joint outcomes.

Understanding human mental models of AI, that is how humans perceive AI systems including their beliefs about the accuracy, capabilities, and trustworthiness of AI. Misaligned or incomplete mental models can lead to overreliance or underutilization of AI.

Designing effective human–AI interactions by optimizing the timing and information shared by AI to avoid cognitive overload while fostering engagement. This includes developing explainable AI systems and ensuring the interaction aligns with human cognitive limitations and workflows.

Addressing these challenges requires the development of robust human–AI collaboration models to enhance our understanding of this phenomenon. Numerous models have been proposed, and their key features are summarized in the following tables. Specifically, Table 1 presents the components of study design related to human–AI decision-making (Lai et al. 2023).

Table 2 outlines the socio-technical framework of Hybrid Intelligence (HI) systems (Hemmer et al. 2021).

Table 1 Components of study design on human–AI decision-making (Lai et al. 2023)

Task characteristics	Domain	Application area (e.g., healthcare, law, education).
	Risk Level	High Stakes (e.g., justice) vs. Low Stakes (e.g., recommendations).
	Expertise Required	Domain knowledge: experts (e.g., doctors) vs. lay users.
	Subjectivity	Objective (e.g., diagnosis) vs. subjective (e.g., preferences).
AI Assistance	Model Type	Deep learning (e.g., CNN), shallow models (e.g., regression), Wizard of Oz.
	Predictions	Binary, multi-class, continuous outputs.
	Information about prediction	Uncertainty estimates.
	Information about models	Model's performance and accuracy.
	User's Interactivity	Feedback, control over inputs, interactive explanations.
Evaluation Metrics	Evaluation metrics with respect to decision task	Efficacy, efficiency, and level of satisfaction.
	Evaluation metrics with respect to AI	Trust scores, over-/under-reliance (bias/aversion), user's understanding of AI, perceived AI fairness.

Table 2 Socio-technical framework of HI systems (Hemmer et al. 2021)

Collaboration characteristics	Interaction Style	Order of displaying predictions to humans, active interactivity, degree of automation.
Task characteristics	Complexity and diversity	High task difficulty increases reliance on AI, data type (text, images, tables), in and out-of-distribution data.
AI characteristics	Quality of explanation	Confidence scores, accuracy, clarity of information presentation.
Human characteristics	Cognitive and Personality Traits	Biases, trust calibration, social skills, conscientiousness, emotional stability, knowledge of AI and interaction techniques.

Table 3 Systematic framework of human–AI synergy in decision-making (Bao et al. 2023)

AI Affordances	Information Collection: Automated data gathering.	E-commerce data analysis, clinical symptom tracking, job applicant behavior logging.
	Information Processing: Pattern recognition and analysis.	Resume screening, personalized shopping tools.
	Predictive Assistance: decision support.	Medical diagnosis, supply chain optimization, house price predictions.
	Explanatory feature: Explainable AI (XAI) makes decision-making processes transparent to foster user's trust and understanding.	Confidence levels, visual reasoning tools.
Human–AI Synergy Patterns	AI-Centered: AI leads in low-uncertainty tasks, with minimal human intervention.	Chatbots, automated customer service.
	Human-Centered: Humans retain control in high-uncertainty tasks, with AI providing supporting explanations.	Ethical decision-making, complex diagnostics.
	Synergy-Centered: Balanced collaboration where humans and AI complement each other's strengths.	Cancer detection, recidivism prediction.
Outcomes of Synergy	Performance: AI improves decision-making accuracy, efficiency, and bias mitigation.	Fairer hiring, efficient clinical assessments.
	Trust: Built through transparency and reliability.	Adoption in healthcare, marketing.
	Explainability: Clarity in AI decisions to promote understanding and reduce cognitive load.	Visual and symbolic explanations, time-sensitive communication.

Finally, Table 3 provides a systematic framework for achieving human–AI synergy in decision-making processes (Bao et al. 2023).

While various gaps are identified by different authors, this discussion focuses on those concerning the human component present in all models, particularly the influence of biases. In human-centered patterns of human–AI decision-making, the "black box" issue of AI and the cognitive limitations of human thinking pose significant challenges to achieving optimal decisions and true synergy (Bao et al. 2023). The focus of this review is primarily on biases, specifically AB.

AB has been hypothesized to negatively impact performance (Hemmer et al. 2021), yet its effects remain

underexplored, particularly within the context of Explainable AI (XAI) techniques, where conflicting findings highlight the need for further investigation. While some studies suggest that explanations can mitigate cognitive biases, others indicate that transparency may exacerbate biases, leading to over-reliance, under-reliance, or misapplication of the explanation (Bertrand et al. 2022). Additionally, given the increasing application of AI in high-stakes contexts, AB also has important legal implications: while human oversight is legislatively emphasized, existing regulations like the GDPR and AI Act do not adequately address AB in hybrid decisions (Ruschemeier and Hondrich 2024).

A further concern arises with the global deployment of AI systems across diverse populations. Emerging evidence shows that cultural, gender, socioeconomic, and regional factors significantly influence how humans rely on AI, whereas many design assumptions continue to reflect dominant Western perspectives (Ge et al. 2024). Cultural norms shape users' expectations of AI systems. For instance, European Americans prioritize control over AI, viewing it as a tool, while Chinese participants value connection and openness to AI's autonomy. African Americans exhibit intermediate preferences, emphasizing both control and relational aspects. These differences are rooted in broader cultural models of self: individualistic societies often favor hierarchical relationships with AI, while collectivist cultures are more open to viewing AI as a collaborative partner (Ge et al. 2024). Such disparities risk embedding cultural biases into AI systems, potentially making certain populations more vulnerable to AB.

Building on the existing models of human–AI decision-making, this review examines the occurrence of AB in human–AI collaboration, with a focus on potential correlations with factors highlighted in these models, including:

- Task characteristics, as risk, required expertise (Lai et al. 2023), and complexity (Hemmer et al. 2021).
- Cognitive and personality traits (Hemmer et al. 2021) and socio-demographic factors (Ge et al. 2024).
- Human mental models of AI (Steyvers and Kumar 2024) and the effects of human perception of explainable and transparent AI suggestions on decision-making (Bao et al. 2023).
- Cognitive overload (Steyvers and Kumar 2024).
- The order of displaying predictions to humans (Hemmer et al. 2021).
- Interplay with other biases and heuristics (Bertrand et al. 2022).

This endeavor aims to provide an integrative framework for this phenomenon.

1.2 Defining automation bias: a cognitive limitation in human–automation interaction

AB is a cognitive phenomenon where humans display an overreliance on automated systems, favoring automated recommendations over their own judgment, even when contradictory and more accurate information is available (Cummings 2004).

The level of automation, ranging from fully automated to primarily manual solutions, influences the four stages of human information processing: information acquisition, information analysis, decision-making/action selection and action implementation. Within this information-processing framework, AB frequently results in two types of errors: errors of commission and errors of omission. Errors of commission occur when decision-makers take inappropriate actions based on erroneous automated suggestions without verifying them against alternative information sources. Errors of omission, on the other hand, arise when decision-makers did not take appropriate action when not informed or alerted by the decision aid (Wickens et al. 2021).

Previous reviews on this topic identified mediators of AB in user factors, task characteristics and environmental conditions. The first user factor is experience, which acts as a moderator, since experience may decrease overreliance, even though AB can still occur among experienced users; then confidence in the user's own decision also decreases reliance on external support, whereas predisposition to trust was considered the most driving factor. On the other hand, task factors such as complexity and workload were also considered to intensify reliance on automation, along with environmental influences as time pressure and cognitive overload that lead to reallocation of attention and increase AB likelihood (Goddard et al. 2012). Interestingly, AB is not limited to multitasking scenarios but also encompasses single-task scenarios. In fact, it is more closely associated with cognitive load than multitasking per se. Cognitive overload is considered the most significant factor inhibiting users from adequately verifying automation accuracy, thereby increasing AB: the harder the verification complexity, that is the complexity of verifying that the automation is performing correctly, the greater the likelihood of undue trust in the system (Lyell and Coiera 2017). Therefore, cognitive overload and trust remain the most crucial factors since, paradoxically, higher automation accuracy can further reinforce user trust, leading to an increased likelihood of AB. Users often fail to critically evaluate automation outputs when they perceive the system to be highly reliable (Lyell and Coiera 2017).

Overall, some aspects of the AB stem from these attentional/working-memory limitations. However, other aspects of AB are rooted in decisional rather than merely attentional factors. Like other decision heuristics and biases, AB

reflects the tendency of humans to minimize cognitive effort, a tendency described by the "cognitive miser" hypothesis (Wickens et al. 2021). For example, *confirmation bias* and related confirmatory heuristics also contribute to inappropriate reliance on technology. Confirmation bias refers to the tendency to seek, interpret, and remember information that confirms existing beliefs. This bias significantly shapes how users interact with AI systems, especially when they are highly confident in their own judgments while having limited understanding of how AI works. As a result, users may over-rely on models that reinforce their own biases or under-rely on models that might challenge or correct those biases (Lu and Yin 2021). Thus, AB does not operate in isolation but interacts with other cognitive biases.

Another important concept closely related to AB is *automation complacency*. The literature remains divided on whether these should be treated as distinct phenomena or as overlapping forms of automation misuse. Some researchers emphasize differences in their cognitive and situational antecedents (Wickens et al. 2015), while others highlight substantial overlap, viewing both as manifestations of shared underlying attentional and decision-making mechanisms (Parasuraman and Manzey 2010).

Following the approach of previous reviews, we will treat these terms interchangeably for simplicity. Additionally, we will examine the drivers of AB in contrast to those of its counterpart, *algorithmic aversion,* given that both are influenced by users' often unaware expectations about AI performance (Jones-Jang and Park 2023). Both over-reliance and under-reliance are antithetical forms of miscalibrated trust. Striking the right balance remains a major challenge, and we still lack a full understanding of why such extremes are so prevalent.

2 Methodology

The search was conducted using the PRISMA 2020 framework (Page et al. 2021) to examine AB in human–AI collaboration. Considering the occurrence of AB in this context, with a focus on potential correlations with human–AI decision-making models' components (cf. 1.1.1), this review seeks to answer the following research questions (RQs):

RQ1: What are the causal or mediator factors of automation bias?

RQ2: Does automation bias co-occur with other cognitive biases?

RQ3: Can explainability or transparency mitigate automation bias?

RQ4: Is AI literacy or an improved mental model necessary to reduce automation bias?

RQ5: Are individual differences observed in the manifestation of automation bias?

2.1 Search strategy and eligibility criteria

A systematic search was performed on SCOPUS, ScienceDirect, PubMed and Google Scholar. Studies were collected on May 2025.

The included studies had to meet the following criteria: (I) published in peer-reviewed journals in the last 10 years (between January 2015 and April 2025), (II) written in English, (III) use quantitative and experimental designs, (IV) focus on automation bias or over-reliance behavior. While the exclusion criteria were as follows: (I) pre-prints, conference proceedings, theses or dissertations, reviews, and qualitative studies; (II) studies centered on AI acceptability or trust calibration without addressing automation bias; (III) articles discussing the technical development of AI or algorithmic biases unrelated to automation bias.

Before starting the review process, the two independent researchers agreed on the procedure and criteria for evaluating the search strategy. This ensured consistency in their evaluations, which ultimately led to a 0.95 Cohen's Kappa agreement index. In cases of disagreement, critical points were discussed and resolved through a collaborative meeting, ensuring an accurate and consistent evaluation.

Applying the PICo (Population, phenomenon of Interest, Context) search strategy (Stern et al. 2014), the search strategy was constructed using Boolean operators to connect key terms and focus on the scope of the inquiry. The final query was: ("professional*") AND ("automation bias" OR "over-reliance" OR "cognitive bias*") AND ("Human-AI collaboration" OR "Human-AI team*" OR "Human-AI decision making" OR "AI assisted" OR "AI aided" OR "XAI" OR "explainab*"). However, we were able to adopt this complete string in Google Scholar only given its flexibility. In Science Direct we adjusted the query for indexing limitations and syntax constraints.: (professional) AND (TITLE-ABS-KEY ("automation bias" OR "over-reliance" OR "cognitive bias")) AND (TITLE-ABS-KEY ("human-AI collaboration" OR "human-AI decision-making" OR "AI-assisted" OR "explainable AI" OR "XAI")). The same for SCOPUS, where the query was simplified due to Boolean length limitations. Also, the year range (2015–2025) was embedded directly in the search query because SCOPUS does not support standalone date filters in its basic or advanced interface: (professional) AND (TITLE-ABS-KEY ("automation bias" OR "overreliance" OR "cognitive bias")) AND (TITLE-ABS-KEY ("human AI collaboration" OR "human AI team" OR "human AI decision making" OR "AI assisted" OR "AI aided" OR "explainable AI" OR "XAI")) AND PUBYEAR > 2014 AND PUBYEAR < 2026.

In PubMed, we used a hybrid query combining MeSH terms and free-text fields to account for both indexed and emerging terminology. The final query was: ("automation bias"[tiab] OR "over-reliance"[tiab])

AND ("artificial intelligence"[MeSH Terms] OR "artificial intelligence"[tiab] OR "AI"[tiab]) AND ("decision making"[MeSH Terms] OR "decision making"[tiab]). The query was pilot-tested against five benchmark studies and yielded 31 results, of which 3 were included after full-text screening; whereas Google Scholar yielded 17 results and SCOPUS one. Once the records were obtained, we exported them to Zotero to detect duplicates. The complete search and screening process is illustrated in Fig. 1 showing PRISMA flow diagram.

In addition to the structured database search, we employed supplementary search techniques to enhance the comprehensiveness of the review. Reference lists of included articles were manually screened to identify additional studies, performing a backward citation searching as recommended by PRISMA guidelines (Page et al. 2021; Rethlefsen et al. 2021). In accordance with PRISMA-S guidelines (Rethlefsen et al. 2021), we also included records identified through personal files. These were reported under 'other sources' and screened using the same eligibility criteria as records identified through structured database searches. Both these sources are reported under "Identification of studies via other methods" in the PRISMA flow diagram.

2.2 Screening and selection

To fulfill the initial screening, titles and abstracts were screened to identify studies relevant to the research question. Articles passing the initial screening were reviewed in full text to confirm their relevance based on the eligibility criteria. Finally, 35 studies were included in review (Table 4).

3 Results

The 35 selected studies were published in 21 different journals spanning multiple disciplines such as cognitive psychology, human factors engineering, human–computer interaction, and healthcare. The most relevant journals for our review were *Scientific Reports* (n = 2), *Cognitive Research: Principles and Implications* (n = 2), *Human Factors* (n = 2), *Artificial Intelligence in Medicine* (n = 2), *International Journal of Human–Computer Studies* (n = 2), and the ACM journals (n = 5). All the studies were published between 2015 and 2025, with 2023 and 2024 being the most prolific years (n = 10; 28.5% each).

In total, the reviewed studies included 19,774 participants. Owing to our inclusion criteria, all the studies employed quantitative methods. The studies primarily

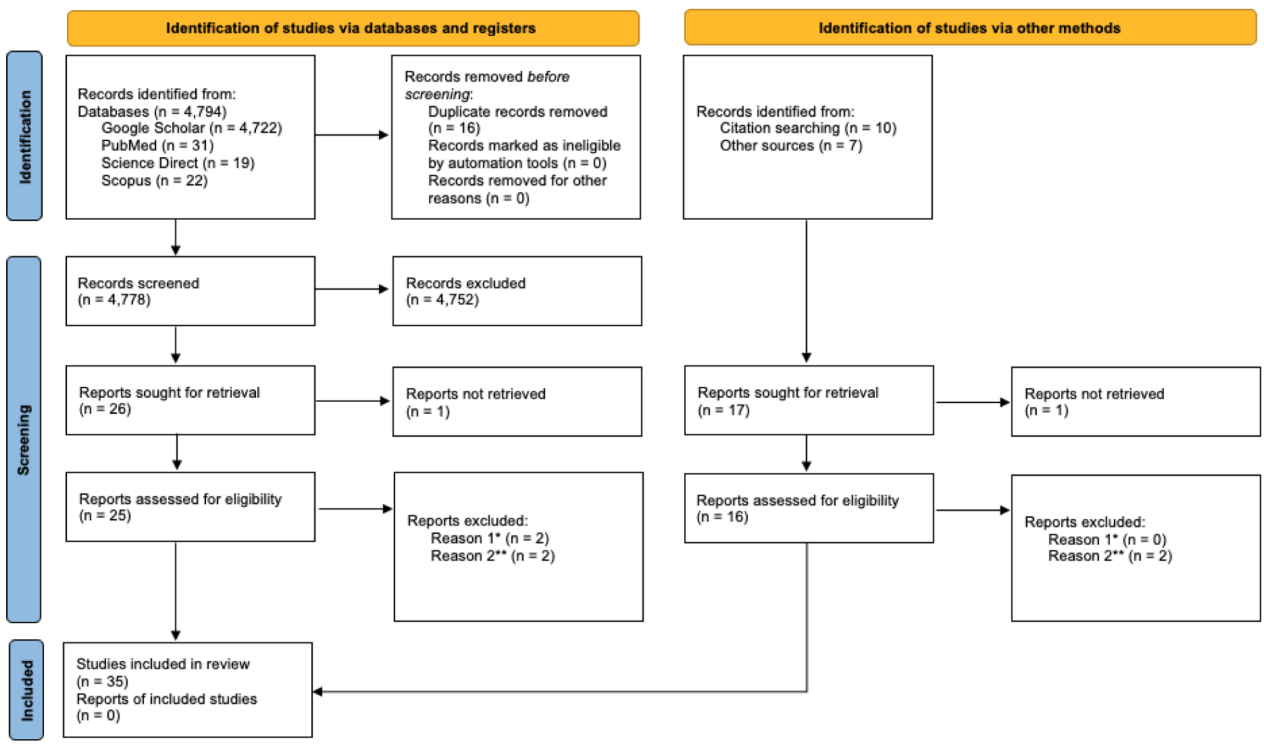

Fig. 1 PRISMA chart. (*Legend:* *studies that examine other biases in human–AI interaction, without addressing the potential priming effects of AI; ** studies that don't offer a sufficient exploration of the underlying cognitive drivers of automation bias, nor do they examine it through behavioral observation)

focused on decision-making tasks involving AI-assisted recommendations. Although displaying heterogeneity in their study-design and measured variables, they consistently employed experimental methodologies focusing on objective performance as the primary dependent variable (e.g., diagnostic accuracy or decision quality, depending on the task domain). Consistent with prior reviews, healthcare emerged as the most frequently studied domain (n = 11; 31.4%). Other investigated domains included finance (n = 3), national security (n = 5), public administration (n = 3), human resources (n = 2), and mental healthcare (n = 2). Across these contexts, agreement with incorrect AI recommendations was the most used operational measure of AB (e.g., Buçinca et al. 2021). AB has also been measured as how often users completely change their answers to match AI recommendations (e.g., Horowitz and Kahn 2024) or as Weight of advice (WOA), which assesses the importance users give to AI recommendations (e.g., Küper et al. 2025).

In addressing AB, the study by Bond et al. (2018) investigated the impact of automated diagnoses (ADs) on ECG interpretation among 30 physicians (15 cardiology fellows, 15 non-fellows). Physicians interpreted ECGs under three conditions: without AD, with a single AD (50% incorrect), and with multiple ADs. Incorrect ADs significantly reduced accuracy (fellows: 84.86–41.66%; non-fellows: 86.38–27.43%), with non-fellows more susceptible to AB. In addition, confidence ratings predicted accuracy in fellows, although both groups reported reduced confidence when incorrect ADs were presented.

Gaube et al. (2021) assessed how the source of diagnostic advice (AI vs. human) influenced perception of advice quality and diagnostic accuracy among radiologists (high-expertise group) and other physicians (lower expertise group). The results revealed that, although radiologists rated AI advice lower in quality, exhibiting algorithmic aversion, they relied on it anyway. Specifically, researchers reported a general tendency of both groups to follow confirmatory heuristics (to human and algorithmic advice) as anchoring effect and confirmation bias, but especially among lower experts. Moreover, AI advice, showing no uncertainty compared to humans, is supposed to make less experts more susceptible to over-trust.

Contrasting with the view of experts as more averse to algorithmic output, the study by Keding and Meissner (2021) examined AI-augmented decision-making among 150 senior executives making R&D investment decisions. AI advice enhanced users' confidence and decision quality, but also increased overestimation of AI's reliability, leading to overreliance due to its perceived objectivity.

Next, involving 27 radiologists with varying experience levels, Dratsch et al. (2023) examined AI-generated BI-RADS scores for mammogram interpretation. The impact of incorrect AI predictions showed accuracy dropping significantly for all groups (unexperienced: 79.7–19.8%, moderately experienced: 81.3–24.8%, highly experienced: 82.3–45.5%). In addition, less experienced radiologists exhibited more commission errors (i.e., accepting incorrect higher-risk categories), whereas errors of omission were similar across all groups. In the same vein, Matzen et al. (2024) found that users had more difficulty detecting errors where AI missed a target rather than false alarms. These findings suggest that AI could support novices perform at levels closer to experienced ones but poses deskilling risks (Dratsch et al. 2023).

Consistent with this, Kim et al. (2025) found that false-positive AI suggestions affected radiologists' diagnoses of cerebral aneurysms. However, the very experienced group still maintained the highest and most stable diagnostic performance. Küper et al. (2025) further investigated reliance behaviors in dermatology and found that medical experience was linked to self-reliance, while the psychological trait of dispositional trust indirectly influenced overreliance on AI via its effect on situational trust.

Drawing on the distinction between dispositional, situational, and learned trust, Duan et al. (2024) found that overreliance tends to occur uniquely with AI, in contrast to human teammates. In the process of trust development, users may inappropriately transfer learned trust from one AI interaction to future systems or contexts. This contrasts with interpersonal trust, where individuals typically do not generalize one person's reliability to another.

Sato et al. (2020) further explored reliance on imperfect automated signaling systems with 70% reliability in high-risk environments. Their participants in the high-risk group showed higher trust levels, particularly under high load, regardless of the actual system reliability. Although task load was high, operators still allocated attention effectively to monitor the system, but this did not result in reduced trust in high-risk condition.

Kücking et al. (2024) analyzed wound assessment in 210 professionals (63.3% nurses, 36.7% physicians) using AI-enabled Clinical Decision Support Systems. Their findings identified several protective factors against AB: diagnostic expertise, wound care training, and female gender. In contrast, a higher perceived benefit of AI was associated with increased susceptibility to AB, while age was not a significant factor. In line with these results, Dikmen and Burns (2022) found that providing domain-specific knowledge helped non-expert users to critically evaluate AI outputs, relying less on incorrect AI predictions.

Then, Klingbeil et al. (2024) employed a trust game mimicking economics study. Participants were assigned roles as proposers (Player A) or responders (Player B) in financial cooperation tasks. Players A had to decide whether to cooperate with Players B based on their decision history, receiving advice labeled as either AI or human expert generated.

Results showed that participants exhibited greater reliance on AI even when it contradicted available contextual information. Additionally, situational trust positively correlated with advice conformity, more than dispositional trust.

In contrast to all these findings, Alon-Barkat and Busuioc (2023) found *selective adherence* to AI suggestions, rather than AB, in employment decisions. In this study, participants were more likely to follow advice aligned with their pre-existing stereotypes (e.g., against ethnic minority indicators like a Moroccan-sounding name): they were more inclined to terminate contracts of Moroccan-sounding teachers when scores were low, despite similar scores for Dutch teachers. In like manner, the studies by Selten et al. (2023) and Bashkirova and Krpan (2024) found no evidence for AB, emphasizing *confirmation bias* as a more dominant cognitive mechanism in AI-assisted decision-making.

Thus far, studies have depicted experts as either more resistant to AI influence or still susceptible to AB. The following studies explore the factors that drive these differing behaviors within expert groups. Additionally, they examine the potential role of explainability in mitigating AB.

On this point, Rezazade Mehrizi et al. (2023) investigated how AI explainability tools (like heatmaps) and attitudinal priming (positive or negative information about AI) influence radiologists' diagnostic decisions when interpreting mammograms. The study revealed that heatmaps improved accuracy only when AI suggestions were correct but tended to promote over-diagnosis when AI suggestions were wrong; likely because radiologists used the heatmaps as visual heuristics, focusing on highlighted areas while neglecting others. Overall, there was no significant difference across all groups to both conditions.

Nourani et al. (2022) further examined how *anchoring effect* influences users' mental models when interacting with intelligent systems. Anchoring bias, in this context, refers to the impact of first impressions, particularly, the order in which users encounter the system's strengths and weaknesses. Encountering system strengths first led to increased reliance and more errors, while encountering weaknesses first reduced errors but caused underestimation of system capabilities. The study also showed significant differences in impression formation based on participants' domain expertise: novices exhibited greater AB, while experts recalibrated trust over time depending on what impression had first. In fact, experts' trust is more sensitive to the order in which the wrong advice is presented. Researchers also explored the role of explanations in addressing this bias and found that explanations improved user's self-confidence but did not mitigate AB caused by first impressions.

Vered et al. (2023) and Cecil et al. (2024) also challenged the prevailing belief that explainability improves decision-making. However, they attribute overreliance on incorrect advice to different factors. Vered et al. (2023) argued that AI explanations reinforced unwarranted trust, while Cecil et al. (2024) showed that complex explanations increased cognitive load, hindering effective processing. As additional EEG evidence supported, increased workload induced by decision difficulty contributed to overreliance by adding cognitive strain (Zhang et al. 2024).

Attempts to address this issue through multiple types of explanation have proven ineffective, as offering varied explanation formats did not significantly improve users' ability to detect incorrect AI recommendations (Naiseh et al. 2023). Among these, simplistic explanations (with limited information) exacerbated this vulnerability the most (Jacobs et al. 2021). Offering trust calibration feedback did not help either (Tatasciore and Loft 2025).

Then, Buçinca et al. (2021) adopted cognitive forcing functions (CFFs) as a debiasing tool. Researchers tested six conditions: three CFF-based (on-demand, update, wait), two simple explainable AI (SXAI) approaches (explanations, uncertainty), and a no-AI baseline. Results showed that when AI predictions were correct, CFFs and SXAI performed similarly. Differently, CFFs significantly reduced overreliance on AI compared to SXAI approaches when the AI's recommendations were incorrect. Also, participants trusted and preferred SXAI systems over CFF-based systems as CFFs were perceived as more cognitively demanding; in fact, subjective trust ratings (preference/acceptability) negatively correlated with objective performance. This highlights a disconnect between perceived and actual benefit of relying on AI, as also reported by Jacobs et al. (2021) and Cao et al. (2023).

The update CFF condition involved participants making an initial decision without access to AI assistance, followed by the presentation of the AI's suggestion, which they could then use to revise their decision. The manipulation of this condition was further examined in subsequent studies. Specifically, Cabitza et al. (2023) examined human–AI collaboration in medical diagnostics, comparing AI-first and human-first decision protocols. AI-first protocols, in which the AI provided recommendations before the human reader made their own evaluation, improved diagnostic accuracy, but also increased the priming effect of AI advice when it was incorrect. Human-first protocols, where humans made initial assessments before receiving AI output, led to either conservatism bias (self-reliance on their initial judgment) or automation complacency (excessive future trust in AI) depending on whether the AI recommendations confirmed or conflicted with their initial evaluations. Both these responses were considered as forms of anchoring effect. Additionally, explanations increased AI trust but did not improve diagnostic accuracy. Likewise, Agudo et al. (2024) observed that participants who received AI suggestions before forming their own judgments were significantly more likely to align with incorrect AI assessments.

Cao et al. (2023) further investigated human reliance on AI in human-first settings under time pressure, manipulated via limited initial observation time (before seeing AI input) and final decision time (after viewing AI suggestions). They found that shorter decision time significantly increased the anchoring effect of AI, while longer observation time encouraged users to trust their own judgment. However, these effects varied by task: in more logically involving tasks, like spatial reasoning, users showed greater resistance to AI suggestions even under time constraints. Consistent with this, Rastogi et al. (2022) demonstrated that time is a useful resource for de-anchoring.

Thereafter, investigating the use of AI dashboards in HR selection, Kupfer et al. (2023) found that higher verification intensity (i.e., thorough review of AI recommendations) improved decision accuracy. Secondly, being informed on the potential AI errors increased verification efforts, but emphasizing responsibility did not significantly affect verification intensity. Lastly, highly aggregated data visualizations on the dashboards reduced verification behaviors although not impairing decision quality.

Finally, Horowitz and Kahn (2024) examined trust and reliance on AI across users' different levels of AI literacy. Individuals with minimal AI background knowledge are more prone to aversion, while those with moderate knowledge are most likely to over-rely on AI, exhibiting the *Dunning–Kruger effect*. This bias makes individuals with average AI background knowledge overestimate their understanding of it, leading to overreliance on automation. Instead, participants with extensive AI knowledge demonstrated the most balanced reliance.

4 Discussion

4.1 Unraveling the influencing factors

In this section, we analyze the findings from the included studies, comparing them with the factors identified in previous reviews by Goddard et al. (2012) and Lyell and Coiera (2017). Goddard et al. (2012) highlighted user's experience and confidence as protective factors against AB, while predisposition to trust emerged as the most influential driver. Additionally, they identified environmental mediators such as workload, task complexity, and time constraints. Lyell and Coiera (2017) further emphasized that task complexity and verification demands contribute to cognitive overload, increasing susceptibility to AB.

4.1.1 Confidence

Jacobs et al. (2021) identified no significant change in confidence across conditions involving correct or incorrect AI recommendations, suggesting that confidence may not protect against AB. Similarly, Bond et al.'s (2018) findings diverge from the conclusions drawn by Goddard et al.'s (2012) review. Specifically, even though experts can still be affected by AB, they suggest that only among expert clinicians confidence and experience are predictors of accuracy and can reduce overreliance. Less experienced clinicians, despite their confidence, still demonstrated greater vulnerability to AB (Bond et al. 2018), showing that confidence alone is not a sufficient protective factor. Additionally, Gaube et al. (2021) established a link between confidence and perceived AI suggestion quality. Physicians' confidence decreased when the advice was inaccurate; in fact, although experienced radiologists tended to show more algorithmic aversion, surprisingly, their expressed aversion did not affect their reliance on the AI-generated advice (Gaube et al. 2021). This does not imply that confidence itself is predictive of accuracy. Instead, it just suggests that when physicians perceive the advice as high quality (i.e., correct), they are more confident in their diagnostic decisions. Meanwhile, Dratsch et al. (2023) further illustrated that confidence and experience reduced only errors of commission. Specifically, AB affected all the groups of radiologists, regardless of their experience, but inexperienced radiologists were more prone to errors of commission, whereas errors of omission were similar across all groups.

By contrast, Cabitza et al.'s (2024) and Küper et al.'s (2025) data continued to support a positive correlation between confidence and accuracy. However, it is also possible that confidence is influenced by the clinician's level of experience (Glick et al. 2022), which has been shown to be more predictive of diagnostic accuracy. In fact, across the studies, there's a broad consensus -most explicitly outlined in the study by Kücking et al. (2024)—that professional experience remains the most critical protective factor against AB. In diagnosing wound healing complications using Clinical Decision Support Systems, participants with better initial diagnostic performance and those with specific training in wound care were less likely to agree with false AI recommendations. This means that although non-specialists benefit from AI assistance, domain-specific education and expertise are still needed to use these tools properly (Kücking et al. 2024).

4.1.2 Trust as state, trait, and learned belief

Trust in technology can be conceptualized either as a state or a trait, each carrying distinct psychological and behavioral implications. Trust as a state is a situational and dynamic response based on contextual factors, such as the perceived reliability or transparency of a technological system. In contrast, trust as a trait, or propensity to trust, reflects a psychological predisposition to trust others, including machines

(Küper et al. 2025). A third dimension is learned trust, which develops over time through repeated interactions, as users integrate past experiences in future expectations and behavior (Duan et al. 2024).

Within this process of trust development, two key aspects shape trust: cognition-based trust, grounded in users' perceived understandability and technical competence of AI, and affect-based trust, based on emotional reactions to the system's behavior (Naiseh et al. 2023; Duan et al. 2024). A further influence is the role of cognitive biases, which can distort perceptions of reliability (Naiseh et al. 2023),

Across the studies, trust consistently emerged as a central determinant of overreliance: Küper et al. (2025) reported that trust accounted for up to 24.1% of the variance of reliant behavior, whether appropriate or not. Similarly, Horowitz and Kahn (2024) described trust as functioning like a threshold: when initial trust in AI is low, users are reluctant to rely on it, regardless of its demonstrated reliability. Once trust is established, reliability cues, such as extensive training, become a dominant factor in the willingness to rely on technology (Horowitz and Kahn 2024). In addition to perceived reliability, Buçinca et al. (2021) also found that people preferred AI systems that were less cognitively demanding, even if those systems did not improve decision quality. These findings align with existing literature on trust in automation, which emphasizes the roles of system complexity, user familiarity, and perceived reliability (see Wickens et al. 2021).

Another noteworthy finding comes from Klingbeil et al. (2024), who challenged the earlier claims by Goddard et al. (2012) regarding the primacy of dispositional trust. Instead, they emphasized the role of situational trust, shaped by the specific context and task. Situational trust positively correlated with reliance on AI or human advice, but propensity to trust correlated negatively, suggesting that while trust influences decision-making, a higher tendency to trust could lead to more cautious behavior (Klingbeil et al. 2024). By contrast, Küper et al. (2025) still argue that dispositional trust is a meaningful antecedent of initial trust, but they also highlight that situational trust acts as a mediating factor, particularly in predicting overreliance.

Expanding on this distinction between dispositional, situational, and learned trust, Duan et al. (2024) outlined a developmental trajectory of trust in AI. Initial trust is shaped by expectations of AI, then it gets fostered or hindered during tasks based on situational factors, such as communication effectiveness and acknowledgment of errors. Over time, it eventually gets consolidated through experience as learned trust. However, this process may lead to miscalibrated trust in AI specifically: unlike interpersonal trust, which is rarely generalized across individuals, users tend to transfer learned trust from one AI system to another, making overreliance more likely.

On the whole, trust can both enhance or impair decision accuracy, especially among inexperienced users, by enabling beneficial or detrimental overreliance (Cabitza et al. 2023). Its impact often depends on the correctness of AI advice, making the outcome of trust, whether appropriate or inappropriate, difficult to predict (Naiseh et al. 2023). Yet, trust alone does not fully account for AB, as some samples tended to over-rely (e.g., Klingbeil et al. 2024), while others under-rely (e.g., Küper et al. 2025), underscoring the need to consider its interaction with other factors, explored in the sections that follow.

4.1.3 Interplay with other biases

Nourani et al. (2022) associated AB with the *anchoring effect*, wherein users overly depend on initial information. Similarly, Gaube et al. (2021) highlighted the role of this confirmatory bias,observing that AI recommendations prime users to seek corroborative evidence, a behavior also observed within traditional decision-making contexts. Interestingly, such confirmatory bias can also contribute to conservativism in human-first protocols, as noted by Cabitza et al. (2023).

To explain, respectively, overreliance and aversion, Klingbeil et al. (2024) proposed the MABA–HABA framework ("Machines Are Better At vs. Humans Are Better At"), suggesting that reliance is based on perceptions of which agent (AI or human) is better suited for the task. While machines are trusted for mechanical tasks and objective calculations, humans are preferred for social, emotional, or ethical judgments. However, this framework does not fully account for reliance on AI in financial decision-making, which involves both dimensions (Klingbeil et al. 2024). Also, it could not account for the occurrence of opposite behaviors for the same domain task that Cabitza et al. (2023) portrayed.

An alternative explanation is that other heuristics come into play. Negative experiences, such as conflicting results, can erode user trust in AI, as explained by the negativity bias and the illusion of validity. Experts, in particular, are more likely to reject AI recommendations due to reliance on their established heuristics and the added pressure of high-stakes settings. These environments often lead them to develop cognitive routes that enable them to make quick and accurate decisions. While their intuition is more refined than non-experts' one, it also makes experts more susceptible to belief perseverance and algorithmic aversion (Bertrand et al. 2022).

In this regard, Nourani et al. (2022) highlight how *ordering bias* impacts knowledgeable users: experts display higher sensitivity to the order of error. Positive initial experiences tend to foster trust and tolerance for future errors, whereas negative first impressions may cause lasting distrust (Nourani et al. 2022). This phenomenon may explain

discrepancies in behavior among individuals with similar levels of expertise, suggesting that personal experiences shape trust dynamics.

In addition to that, other results suggest that a further layer of knowledge beyond domain expertise may be involved. The study by Horowitz and Kahn (2024) revealed the non-linear relationship between AI background knowledge, trust, and AB. They illustrated this dynamic through a function representing the rate at which participants switched their decisions after receiving AI recommendations. Individuals with minimal AI background knowledge were more prone to aversion. Those with moderate levels of background knowledge were relatively over-reliant on AI, while participants with the highest levels of background demonstrated appropriate reliance. This pattern suggests that individuals with a moderate AI background are affected by the *Dunning–Kruger effect*, a cognitive bias in which those with superficial knowledge become overconfident in that knowledge.

Drawing on all these findings, we hypothesize that aversion and over-reliance are shaped by the intersection of two types of knowledge: domain expertise and AI-specific background knowledge, each bringing its own heuristics and biases into the decision-making process.

4.1.4 Transparency and human mental model of AI

Human–AI collaboration models emphasize the importance of users' mental models of AI in decision-making processes (Bao et al. 2023; Steyvers and Kumar 2024). One proposed strategy to mitigate AB is through transparency, particularly by providing uncertainty estimates, which may help users recognize the limitations of AI systems (Bond et al. 2018; Gaube et al. 2021; Dratsch et al. 2023). Uncertainty estimates are intended to help users identify when the AI might be wrong, aiming at misattributed trust. However, as we mentioned, some biases are exacerbated by explanations, leading to over-reliance, under-reliance, or misapplication of the explanation. Over-reliance occurs when users overly trust AI after being provided with any explanation (mere exposure effect). Also, longer explanations seem more plausible (completeness bias), while visual explanations may amplify misplaced trust due to confirmation bias (Bertrand et al. 2022).

Notably, Nourani et al. (2022) highlight how the anchoring effect of first impressions influences the formation of mental models, but explanations failed to mitigate the anchoring in both positive-first and negative-first impression scenarios. In an effort to reduce over-reliance, Buçinca et al. (2021) adopted CFFs, a metacognitive debiasing technique aimed at promoting analytical thinking. Their findings showed that CFFs significantly reduced reliance on incorrect AI recommendations compared to simple XAI approaches. Nonetheless, users still preferred simpler XAI systems, viewing CFFs as more cognitively demanding. Although coming at the cost of user preference, this suggests that increased mental demand can foster critical engagement. Indeed, explanation-only approaches influence overreliance through either attitude-related or heuristic pathways. Closely, Vered et al. (2023) argued that explanations provided by XAI systems can reinforce unwarranted trust, as users may interpret these explanations as endorsements of the AI's trustworthiness rather than tools for critical evaluation. On the heuristic side, Cecil et al. (2024) found that complex explanations increase cognitive load, making it harder for users to process information effectively. Similarly, Jacobs et al. (2021) argued that overly simple explanations, limited in informational content, are also particularly misleading when the AI is incorrect, compared to more detailed explanation types.

Conversely, Vasconcelos et al. (2023) offer a different view, suggesting that explanations can reduce overreliance, but only under specific conditions. Drawing on a cost–benefit framework, they argue that users assign higher utility to easier explanations in harder tasks. This creates an apparent contradiction: while Buçinca et al. (2021) suggest that complexity fosters critical engagement, Vasconcelos et al. (2023) argue that simplicity enhances explanations effectiveness. Rather than a direct conflict, these findings reflect a broader Simplicity vs. Detail tension in XAI design, a theme we will explore further in Sect. 5.1.

Lastly, Kupfer et al. (2023) found that when decision-makers were made aware of potential system errors, they exhibited greater verification intensity. However, as Rezazade Mehrizi et al. (2023) caution, such interventions may be ineffective if users lack familiarity with AI tools, underscoring the need for both awareness and experience to ensure optimal decision support.

4.1.5 Personality traits and individual differences

Kupfer et al. (2023) observed high standard deviations for verification intensity indicators within experimental groups: participants lower-scored on verification intensity indicators displayed higher AB. Hence, they suggested to start research in individual differences. In this regard, Buçinca et al. (2021) revealed the impact of individual differences in cognitive motivation (Need For Cognition, NFC), which expresses how an individual enjoys engaging in cognitively demanding activities. High-NFC participants benefited more from CFFs, showing greater improvements in decision-making. Moreover, Nourani et al. (2022) showed how different levels of user's domain expertise cause them to form different understandings of explanations, as well as user's past experiences, their conceptions about how AI works, and the view of AI held by their social context.

4.1.6 Task characteristics: complexity and risk

Occurrences of overreliance may also be understood as strategies to reduce cognitive load, rather than as mere manifestations of misplaced trust (Alami et al. 2025). Both cognitive load and task complexity have been confirmed to influence susceptibility to AB (Cecil et al. 2024; Zhang et al. 2024). These findings support the significance of verification complexity, as previously highlighted by Lyell and Coiera (2017). As shown by Kupfer et al. (2023), higher verification intensity helps counteract errors.

Notably, engaging in explanations may be conceived as part of the task; in fact, within the scope of verification intensity, the format of data visualization in AI dashboards influenced user behavior. Highly aggregated visualizations (e.g., a single AI-generated matching score) did not significantly affect decision quality compared to less aggregated formats (e.g., separate scores for education, skills, and personality). However, the highly aggregated formats were found to reduce verification intensity, as users spent less time interacting with the initial levels of the dashboard. This may have promoted more heuristic-based decisions, potentially increasing the risk of AB. These findings suggest that explanations themselves can be used as a heuristic (Rezazade Mehrizi et al. 2023), resulting less effective for high workload contexts as people can't easily detect AI mistakes (Zhang et al. 2024). In any case, increased verification effort tends to correlate with better decision outcomes; however, the conditions that facilitate such engagement must also be considered. In cognitively demanding tasks, additional informational detail can overwhelm users, potentially hindering rather than enhancing decision quality. This issue will be discussed in greater depth in Sect. 5.2.

Risk has also been examined (Sato et al. 2020) but, based on our analysis, it does not qualify as a significant environmental factor affecting AB. Lay users (e.g., consumers) have likewise shown a tendency to over-rely on algorithms even for low-stakes tasks, such as personal preference decisions (Banker and Khetani 2019), as evidenced by the extensive literature on recommender systems (see Adomavicius et al. 2013). Additionally, accountability, which has been proposed as an intervention to mitigate AB (Goddard et al. 2012), lacks empirical support. Kupfer et al. (2023) found no significant effect of responsibility measures in reducing AB. This finding aligns with Lyell and Coiera's (2017) consideration that accountability alone is an ineffective intervention. Instead, efforts should focus on addressing the cognitive load generated by the verification process itself (Fig. 2).

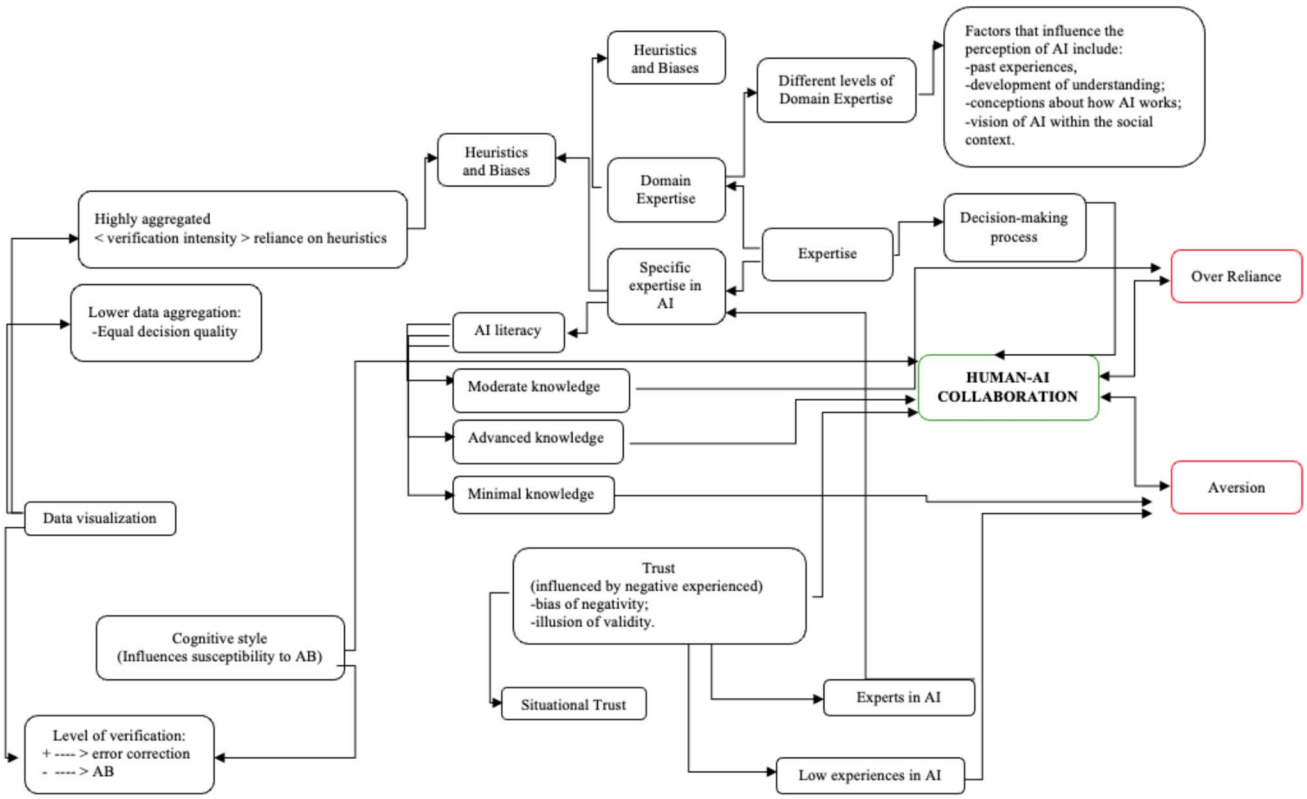

Fig. 2 Visual model illustrating the interaction of various features within human–AI collaboration contexts

Table 4 Articles reviewed after initial screening and eligibility assessment

	Authors	Domain	Population	Task	Type of assistance
1	Bond et al. (2018)	Healthcare	N = 30	ECG interpretations	Automated Diagnosis
2	Sato et al. (2020)	Defense	N = 40	Tracking and system monitoring tasks concurrently	Alerted-monitoring system
3	Buçinca et al. (2021)	Nutrition	N = 199	Ingredient substitution	Cognitive Forcing Functions (CFFs) or SXAI
4	Gaube et al. (2021)	Healthcare	N = 138 radiologists, N = 127 physicians (IM/EM)	Chest X-ray diagnoses	AI-generated diagnostic advice
5	Keding and Meissner (2021)	Finance	N = 150	R&D investment decisions	AI-based advisory system
6	Jacobs et al. (2021)	Mental Healthcare	N = 220 psychiatrists	Antidepressant prescription	none, placebo, feature-based, heuristic-based explanations
7	Dikmen and Burns (2022)	Finance	N = 40 University students	Loan evaluation	AI-only vs AI + Domain Knowledge
8	Glick et al. (2022)	Healthcare	N = 41 Dental students	Radiographic diagnosis of Furcation Involvement	CNN model
9	Nourani et al. (2022)	Mixed Domains	N = 116	Interactive AI applications	AI-generated explanations
10	Rastogi et al. (2022)	Education	Study 1 (N = 47), Study 2 (N = 479) Turkers from Amazon MT platform	Student performance prediction tasks	AI predictions
11	Alon-Barkat and Busuioc (2023)	Public Administration	Study 1 (N = 605), Study 2 (N = 904), Study 3 (N = 1345)	Teacher employment decisions	Algorithmic and human recommendations
12	Cabitza et al. (2023)	Healthcare	N = 12 radiologists N = 44 cardiologists	ECG and MRI Diagnostic tasks	Human-first vs AI-first protocols
13	Cao et al. (2023)	General-purpose decision-making	N = 40	Spatial Reasoning task and Count Estimation task	AI suggestion
14	Dratsch et al. (2023)	Healthcare	N = 27 radiologists	Mammography diagnosis	AI-generated BI-RADS suggestions
15	Kupfer et al. (2023)	HR	N = 93	Candidate selection	AI dashboard with visualizations
16	Naiseh et al. (2023)	Healthcare	N = 41	Chemotherapy prescription screening	Local, Global, Counterfactual, or Example-based Explanations
17	Rezazade Mehrizi et al. (2023)	Healthcare	N = 92 radiologists	Mammography diagnosis	XAI outputs (heatmaps and numerical scores)
18	Selten et al. (2023)	Public Administration	N = 124 police officers	Evaluating burglary, ATM robbery or stabbing incident scenarios	Congruent or incongruent recommendations
19	Vasconcelos et al. (2023)	General-purpose decision-making	N = 731	Maze-solving task	XAI
20	Vered et al. (2023)	Mixed Domains	Study 1 (N = 296), Study 2 (N = 424)	Decision-making with XAI	Local vs Global Explanations
21	Agudo et al. (2024)	Public Administration	Study 1 (N = 80), Study 2 (N = 160)	Assessment of a defendant's guilt	AI probabilistic recommendations
22	Bashkirova and Krpan (2024)	Mental Healthcare	N = 114	Mental health diagnoses	AI-based Triage recommendations, congruent vs incongruent
23	Cabitza et al. (2024)	Healthcare	N = 16 orthopedists	Detecting vertebral fractures in X-rays	Pro-hoc explanations
24	Cecil et al. (2024)	HR	N = 1403	Personnel selection	AI-based recommendation

Table 4 (continued)

	Authors	Domain	Population	Task	Type of assistance
25	Duan et al. (2024)	Defense	N = 45	Reconnaissance photographs task	Wizard of Oz AI agent
26	Horowitz and Kahn (2024)	Defense	N = 9,000	Airplane classification task	AI-based recommendations
27	Klingbeil et al. (2024)	Finance	N = 319	Financial cooperation tasks	AI and human advice
28	Kücking et al. (2024)	Healthcare	N = 210	Wound healing diagnosis	Clinical Decision Support System
29	Matzen et al. (2024)	General-purpose decision-making	Study 1 (N = 37), Study 2 (N = 208), Study 3 (N = 587), Study 4 (N = 303), Study 5 (N = 300)	Visual object detection task	CAD systems
30	Zhang et al. (2024)	General-purpose decision-making	N = 23	Image classification	XAI
31	Alami et al. (2025)	Defense	N = 30	UAV command and control simulation	Alert, level 1 XAI, and automated rerouting features
32	Gegoff et al. (2025)	Defense	N = 142	Simulated uninhabited vehicle (UV) management task	Decision recommendation system
33	Kim et al. (2025)	Healthcare	N = 9 radiologists	Cerebral aneurysm diagnosis using TOF-MRA readings	AI-based CAD system for cerebral aneurysm detection
34	Küper et al. (2025)	Healthcare	N = 223 dermatologists	Skin lesion classification	Clinical Decision Support System
35	Tatasciore and Loft (2025)	Defense	N = 160 under graduated students	Uninhabited Vehicles management missions	Transparent automated recommender and Trust calibration feedback

4.2 Answering our RQs

4.2.1 RQ1: What are the causal or mediator factors of automation bias?

Across studies, trust and task complexity are confirmed as predominant factors in AB. Initial trust in AI, coupled with perception of systems reliability, significantly influences reliance on automation (Buçinca et al. 2021; Duan et al. 2024; Horowitz and Kahn 2024). Several factors shape this perceived reliability, including perceived benefit (Kücking et al. 2024), usefulness (Jacobs et al. 2021; Cao et al. 2023), competence (Naiseh et al. 2023), transparency (Vered et al. 2023), absence of uncertainty (Gaube et al. 2021), and the apparent objectivity of AI recommendations (Keding & Meissner 2021). Overreliance is further reinforced by high workload (Vered et al. 2023; Zhang et al. 2024; Alami et al. 2025), time pressure (Cao et al. 2023), and task complexity (Bond et al. 2018; Kupfer et al. 2023; Cecil et al. 2024). Collectively, these elements contribute to situational trust, which correlates with advice conformity (Klingbeil et al. 2024, Küper et al. 2025).

Another mediator factor is the sequence in which advice is given. Positive first impressions can foster excessive trust in automation (Nourani et al. 2022; Vered et al. 2023). The timing of explanations also matters: AI-first protocols may increase user compliance with AI recommendations (Cabitza et al. 2023; Agudo et al. 2024). Alternatively, human-first protocols can lead to automation complacency when the AI suggestions align with the human's initial evaluation (Cabitza et al. 2023).

At the same time, several moderator factors have been identified that mitigate the likelihood of AB: professional experience (Rezazade Mehrizi et al. 2023; Kücking et al. 2024; Kim et al. 2025) and domain expertise (Nourani et al. 2022; Dikmen and Burns 2022) are the most protective. Furthermore, advanced AI background knowledge exhibits more calibrated trust (Horowitz and Kahn 2024). Finally, verification-related cognitive engagement serves as a critical debiasing mechanism (Kupfer et al. 2023).

4.2.2 RQ2: Does automation bias co-occur with other cognitive biases?

AB frequently interacts with other cognitive biases, particularly the anchoring effect, where initial AI recommendations disproportionately shape subsequent decisions. This bias leads decision-makers to seek evidence that confirms the AI initial suggestion (Gaube et al. 2021; Nourani et al. 2022), especially when incorrect AI guidance is provided before they form their own judgment (Agudo et al. 2024). Otherwise, it can also foster automation complacency when

the AI's recommendation, provided as a second opinion, aligns with users' initial judgment (Cabitza et al. 2023).

AB has also been linked to the Dunning–Kruger effect, where individuals with limited AI knowledge overestimate their understanding of it, which makes them more prone to overreliance on automation (Horowitz and Kahn 2024).

4.2.3 RQ3: Can explainability or transparency mitigate automation bias?

Transparency mechanisms, such as uncertainty estimates, have been proposed to counteract overreliance by highlighting system limitations (Bond et al. 2018; Gaube et al. 2021; Dratsch et al. 2023). However, current methods implemented through XAI have often shown that explanations themselves can be used as heuristics (Rezazade Mehrizi et al. 2023) and reinforce trust in AI systems, exacerbating AB when the advice provided is incorrect (Cabitza et al. 2023; Vered et al. 2023; Cecil et al. 2024). Multiple explanation types (Naiseh et al. 2023) and trust calibration feedback (Tatasciore and Loft 2025) are proven not to help users recognize incorrect AI recommendations as well.

4.2.4 RQ4: Is AI literacy or an improved mental model necessary to reduce automation bias?

Developing a better understanding of AI capabilities and limitations has been shown to foster better decision-making and reduce AB (Kupfer et al. 2023). Higher levels of AI literacy enable appropriate reliance on automation (Horowitz and Kahn 2024).

4.2.5 RQ5: Are individual differences observed in the manifestation of automation bias?

Individual differences significantly influence the manifestation of AB. Professional experience and domain expertise play a crucial role, with experts generally less prone to AB, although they can still be affected under conditions of high cognitive load (Bond et al. 2018; Nourani et al. 2022; Dratsch et al. 2023). Personality traits, particularly need for cognition, have been shown to enhance critical engagement and reduce susceptibility to AB (Buçinca et al. 2021). Additionally, familiarity with AI, conceptions about how it works, and prior exposure to it are key factors influencing how individuals interact with and rely on AI systems (Nourani et al. 2022; Horowitz and Kahn 2024).

Demographic factors, such as gender, have also been found to influence AB, with female participants exhibiting lower susceptibility, whereas age appears to have a neutral effect (Kücking et al. 2024).

4.3 Automation bias and missing evidence

Our research has also led to a subset of studies (Alon-Barkat and Busuioc 2023; Selten et al. 2023; Bashkirova and Krpan 2024) that challenge the prevailing belief of AB as a distinct phenomenon in human–AI interaction. They suggest that expert participants are more likely to accept AI recommendations that align with their prior judgments while remaining skeptical of those that do not. In summary, these studies argue that *confirmation bias* primarily drives experts' acceptance or reluctance of AI recommendations.

In response to these findings, it is worth noting that also Cabitza et al. (2023) observed that in human-first protocols either conservatism or automation complacency (both due to confirmatory heuristics) can occur, depending on whether AI recommendations align with users' initial assessments. Furthermore, the three studies in question fail to account for cases where participants rely on incorrect AI recommendations despite their expressed reluctance, as highlighted by Gaube et al. (2021). Additionally, these results might be better understood in the context of the developmental trust calibration trajectory proposed by Horowitz and Kahn (2024) and Duan et al. (2024). For instance, Bashkirova and Krpan (2024) themselves acknowledge that mental health practitioners, characteristically, tend to exhibit a greater aversion to adopting AI technologies in their profession.

5 Implications for AI systems design

While XAI and transparency mechanisms are widely proposed as solutions to improve human–AI interaction, the empirical findings thus far presented suggest that current approaches may unintentionally amplify AB by fostering misplaced trust. Key challenges lie in AI literacy, whose lack can limit user's ability to effectively collaborate with systems (Long and Magerko 2020), and explanation design itself. Merely making AI explainable or transparent is not enough: explanations must actually help users recognize uncertainty or errors (Zhang et al. 2020).

Building on the findings from the reviewed studies, this section explores practical implications for a user interface design that supports human professionals in human–AI collaboration. These proposals seek to solve the three main challenges in hybrid decision-making (cf. 1.1.1; Steyvers and Kumar 2024).

5.1 Balancing simplicity and detail in explanations

Explanations appear to have limited impact on overreliance when either the task is difficult and the explanation is complex, or when the task is easy and the explanation is simple (Vasconcelos et al. 2023). For high-stakes contexts, Kupfer

et al. (2023) highlight the need for explanations that strike a balance between parsimony and detail. Oversimplification may reinforce AB by masking uncertainty and complexity, while overly complex explanations increase cognitive load, reducing users' ability to evaluate AI outputs effectively.

In this regard, Vered et al. (2023) examined the effects of two types of explanations, local vs. global, on AI-assisted decision-making. Local explanations clarify why an AI model made a specific recommendation, whereas global explanations provide a broader understanding of the inner workings of a system. Their findings suggest that an effective XAI paradigm must balance acceptability (building trust) with skepticism (encouraging critical evaluation). Notably, participants exhibited stronger AB when exposed to local explanations following positive first impressions. However, local explanations also improved focus and attention, illustrating the nuanced and apparently contradictory role of explanation design. Extending this work, Naiseh et al. (2023) compared four explanation types: local, example-based, counterfactual, and global. They found that counterfactual and example-based explanations were especially effective in fostering cognition-based trust, which in turn led to greater engagement with the XAI interface. These findings highlight the critical importance of *understandability* in explanation design. Supporting this, Cabitza et al. (2024) demonstrated that presenting physicians with example-based explanations (i.e., similar past cases, not algorithmic predictions) helps stimulate reflection without replacing human judgment.

To further support effective engagement, Naiseh et al. (2023) argue that interface design should incorporate visual analytics, enabling interactive and interpretable explanations that make AI outputs more accessible and cognitively manageable for users.

5.2 Promoting verification engagement

Closely linked to the previous point is the need to balance the level of engagement required for effective verification with the cognitive load it imposes on users. As highlighted by Lyell and Coiera (2017), higher verification intensity is essential for mitigating AB, yet it simultaneously increases users' cognitive demands.

Buçinca et al. (2021) found that systems employing CFFs were perceived as more mentally demanding, but nevertheless led to improved objective decision performance. This suggests that a certain level of cognitive effort, though potentially reducing user preference, can enhance critical engagement and reduce overreliance on AI. Additionally, providing truly informative explanations beyond recommendations can reduce overreliance and promote more independent judgment (Gegoff et al. 2025; Küper et al. 2025). However, explanation complexity presents a clear trade-off: while more detailed explanations may promote deeper cognitive processing, excessive complexity can overwhelm users, leading to cognitive overload (Cecil et al. 2024; Zhang et al. 2024; Alami et al. 2025).

These findings underscore the role of human cognitive motivation in mediating the effectiveness of XAI solutions. Therefore, beyond optimizing explanation attributes, such as soundness, completeness, faithfulness, sensitivity, and complexity, XAI research must also focus on designing systems that encourage users to actively engage with explanations (Buçinca et al. 2021). To support this, additional incentives for accurate decision-making may help compensate for low intrinsic motivation among users (Vasconcelos et al. 2023). Moreover, effective XAI design should also consider environmental variables, such as workload, time pressure, or competing priorities, and promote explanation use without significantly increasing cognitive burden (Rastogi et al. 2022, Cao et al. 2023, Alami et al. 2025). Optimal time allocation policies may help humans better de-anchor from the AI.

These considerations lead to the next discussion point.

5.3 Adaptive and personalized explanation design

Buçinca et al. (2021) farther argue that XAI should not assume all users will actively engage with explanations by default. Instead, explanation design should be adaptive, catering to individual differences in cognitive engagement. Moreover, XAI systems should support less experienced users with high dispositional trust by helping them recognize potential AI errors, while encouraging experienced professionals to critically engage with system recommendations (Küper et al. 2025). To achieve this, personalized explanations need to be tailored to user's domain expertise, prior experience, and existing mental model of AI (Nourani et al. 2022). A one-size-fits-all approach to explainability may fail to prevent AB (Naiseh et al. 2023), particularly for users with low AI literacy or domain-specific expertise.

5.4 Guided training and mental model formation

To counteract AB, Nourani et al. (2022) propose guided training sessions that expose users to both AI strengths and limitations before deployment. These sessions aim to foster more accurate mental models by presenting users with examples of both correct and erroneous outputs. This approach aligns with broader findings emphasizing the importance of informing users about how first impressions shape AI trust (Kupfer et al. 2023; Vered et al. 2023).

However, for such interventions to be effective, users also require a certain level of familiarity with AI tools (Rezazade Mehrizi et al. 2023), highlighting the need for training that incorporates both awareness and hands-on experience. As

Naiseh et al. (2023) emphasize, users should be familiarized with system explanations, e.g., through usage scenarios or tutorials, since understandability is a critical driver of user engagement. Additionally, it is recommended to integrate multiple explanation types during task performance to calibrate trust dynamically (Nourani et al. 2022), as having alternatives supports human reasoning (Naiseh et al. 2023) and analogical thinking (Cabitza et al. 2024).

5.5 Human-centered AI: rethinking the oversight model and future directions

Existing regulations often rely on the broad notion of human oversight as a catch-all solution, overlooking how heuristics and AB may persist within it (Ruschemeier and Hondrich 2024). In response, Agudo et al. (2024) propose a fundamental shift in AI-assisted decision-making frameworks. Rather than positioning humans as supervisors verifying AI decisions, they argue that a more effective strategy is to let humans make primary decisions while using AI as a second-opinion system, alerting users to potential errors rather than leading the decision-making process. This ergonomic adjustment may reduce overreliance on AI, preserving human autonomy while still leveraging AI's analytical power.

This human-first protocol, also explored by Cabitza et al. (2023) and Cao et al. (2023), corresponds to one of the three CFFs strategies tested by Buçinca et al. (2021), which, as previously discussed, show demonstrable debiasing effects. Building on this direction, Cabitza et al. (2024) introduce the concept of *Frictional AI*, which involves the use of "cognitively non-invasive" tools that encourage oversight, caution, and responsible commitment in decision-making. A key design innovation in this space is the *pro-hoc* approach, which contrasts with conventional "post-hoc" explainability. In post hoc systems, the AI first offers a prediction or classification and then explains its rationale. In contrast, pro-hoc systems withhold interpreting the case and instead present analogous prior cases, both supportive and contradictory, in relation to the user's initial assessment. This encourages analogical reasoning and reflective evaluation without displacing human agency.

In conclusion, as increased verification effort has been shown to reduce complacency toward AI misrecommendations, future XAI research should explore mechanisms for real-time trust calibration, develop explainability strategies tailored to diverse cognitive profiles (including variations in domain expertise, cognitive motivation, and reasoning styles), and design interactive training sessions that help users build accurate mental models of AI systems. Additionally, incorporating feedback mechanisms that allow users to flag suspected errors and provide input can enhance systems accuracy over time. Further directions are also needed to investigate frictional human-first protocols, which may introduce deliberate cognitive effort to promote more reflective decision-making and prevent professionals deskilling risks.

Ultimately, the goal of explainability should not merely be to increase transparency, but to ensure that explanations actively foster critical thinking, support informed decision-making, and counteract cognitive biases, rather than reinforcing them.

6 Conclusion

According to the complementarity principle, humans collaborating with AI systems should make better decisions than either working alone. However, AB remains a critical challenge in human–AI interaction, particularly as AI systems are increasingly integrated into high-stakes domains. This aimed to clarify the concept of inappropriate reliance, illustrating the complex interplay of factors that contribute to AB. These include other cognitive biases, such as anchoring and the Dunning–Kruger effect, along with dispositional, situational, and learned trust, task complexity, and individual differences in expertise and personality traits. We examined this issue through the lens of trust calibration, a fundamental challenge in human–automation interaction. This approach allowed us to assess not only what factors influence trust, but also the extent to which they contribute to misuse (i.e., AB), or its counterpart, disuse (i.e., algorithmic aversion).

Turning to practical implications, owing to extensive evidence demonstrates that explanations alone are often insufficient to improve decision accuracy or mitigate AB, user *engagement* emerges as the most feasible and impactful point of intervention. Thus, effective debiasing strategies in XAI design should extend beyond transparency alone. Features such as *understandability* and *adaptiveness* have emerged as critical components, as they promote user engagement and verification behavior while accounting for associated cognitive costs. Together, these features contribute to reducing cognitive overload and ensuring a better fit with users' cognitive styles and the demands of their task environments.

However, the present study presents several limitations that should be acknowledged. The inclusion criteria constrained the analysis to peer-reviewed journal articles published between 2015 and 2025, potentially excluding relevant earlier works or cutting-edge insights available in pre-print formats. The restriction to English-language studies may have introduced a linguistic bias, limiting the representation of findings from non-English-speaking regions. Furthermore, the exclusive focus on experimental designs precluded the inclusion of qualitative or mixed-methods research, which could offer richer, contextualized perspectives on AB.

The exclusion criteria further narrowed the scope, omitting conference proceedings, theses, dissertations, and reviews, which may contain preliminary or comprehensive overviews of AB research. Similarly, studies addressing related constructs, such as AI acceptability or trust calibration, were excluded unless directly linked to AB, which may have limited the exploration of its interconnections with broader phenomena. Additionally, research centered on technical AI development or algorithmic biases unrelated to AB was not included, potentially restricting insights into the contextual and technical underpinnings of automation reliance. A further limitation lies in the search method: the inclusion of additional sources beyond databases highlights the limitations of our search strategy. Despite being rigorously structured, the query lacked sensitivity and failed to capture several relevant studies due to indexing gaps and variations in terminology. Lastly, despite our intention to examine cultural and socio-demographic factors, this review was unable to address them adequately, largely due to the current lack of studies focusing on cross-cultural comparisons. The only demographic factors identified were gender and age.

In light of the limitations identified, future research will endeavor to adopt a more integrative and methodologically diverse approach, incorporating qualitative and mixed-methods designs to enrich the understanding of the multifaceted nature of AB. Furthermore, greater emphasis will be placed on cross-cultural comparative studies to systematically investigate the role of cultural and socio-demographic variables, which have thus far remained largely unexplored. Such efforts will be crucial to developing a more nuanced, context-sensitive, and globally generalizable conceptualization of the phenomenon. In addition, expanding the scope to include non-English publications and integrating grey literature, such as conference proceedings and pre-prints, could capture emerging trends and innovative perspectives. Future work should explore the intersections of AB with other constructs like acceptability (i.e., humans' willingness to cooperate) and appropriate reliance, as well with AI systems ability to detect over or under-trust (see Okamara and Yamada 2020), particularly within the context of XAI, to deepen our understanding of this critical issue. Another interesting aspect could be the effect of the AI assistance type (whether deep learning, shallow models, or Wizard of Oz) on user's reliance behavior (see Lai et al. 2023). This socio-technical approach could explore AB within a truly integrated and complete framework.

Author contributions Conceptualization: GR and DC; methodology: GR and DC; formal analysis and investigation: GR; writing—original draft preparation: GR; writing—review and editing: GR and DC; resources: GR; supervision: DC. All authors contributed to the article and approved the submitted version.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests

References

Adomavicius G, Bockstedt JC, Curley SP, Zhang J (2013) Do recommender systems manipulate consumer preferences? A study of anchoring effects. Inf Syst Res 24(4):956–975. https://doi.org/10.1287/isre.2013.0497

Agudo U, Liberal KG, Arrese M, Matute H (2024) The impact of AI errors in a human-in-the-loop process. Cogn Res Princ Implic 9:1. https://doi.org/10.1186/s41235-023-00529-3

Alami J, El Iskandarani M, Riggs SL (2025) The effect of workload and task priority on multitasking performance and reliance on level 1 explainable AI (XAI) use. Hum Fact. https://doi.org/10.1177/00187208251323478

Alon-Barkat S, Busuioc M (2023) Human–AI interactions in public sector decision making: "automation bias" and "selective adherence" to algorithmic advice. J Public Admin Res Theory 33(1):153–169. https://doi.org/10.1093/jopart/muac007

Banker S, Khetani S (2019) Algorithm overdependence: how the use of algorithmic recommendation systems can increase risks to consumer well-being. J Public Policy Mark 38(4):500–515. https://doi.org/10.1177/0743915619858057

Bao Y, Gong W, Yang K (2023) A literature review of human–AI synergy in decision making: from the perspective of affordance actualization theory. Systems 11:442. https://doi.org/10.3390/systems11090442

Bashkirova A, Krpan D (2024) Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Comput Hum Behav Artif Hum 2(1):100066. https://doi.org/10.1016/j.chbah.2024.100066

Bertrand A, Belloum R, Eagan J, Maxwell W (2022) How cognitive biases affect XAI-assisted decision-making: a systematic review. In: AAAI/ACM conference on artificial intelligence, ethics, and society, Aug 2022, Oxford, United Kingdom. https://doi.org/10.1145/3514094.3534164 **(hal-03684457)**

Bond RR, Novotny T, Andrsova I, Koc L, Sisakova M, Finlay D, Guldenring D, McLaughlin J, Peace A, McGilligan V, Leslie SJ, Wang H, Malik M (2018) Automation bias in medicine: the influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. J Electrocardiol 51(6S):S6–S11. https://doi.org/10.1016/j.jelectrocard.2018.08.007

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think. Proc ACM Hum Comput Interact 5:1–21. https://doi.org/10.1145/3449287

Cabitza F, Campagner A, Ronzio L, Cameli M, Mandoli GE, Pastore MC, Sconfienza LM, Folgado D, Barandas M, Gamboa H (2023) Rams, hounds and white boxes: investigating human–AI collaboration protocols in medical diagnosis. Artif Intell Med 138:102506. https://doi.org/10.1016/j.artmed.2023.102506

Cabitza F, Natali C, Famiglini L, Campagner A, Caccavella V, Gallazzi E (2024) Never tell me the odds: investigating pro-hoc explanations in medical decision making. Artif Intell Med 150:102819. https://doi.org/10.1016/j.artmed.2024.102819

Cao S, Gomez C, Huang CM (2023) How time pressure in different phases of Decision-Making influences Human-AI collaboration. Proceedings of the ACM on Human-computer Interaction 7(CSCW2):1–26

Cecil J, Lermer E, Hudecek MFC, Sauer J, Gaube S (2024) Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. Sci Rep 14(1):9736. https://doi.org/10.1038/s41598-024-60220-5

Cummings ML (2004) Automation bias in intelligent time critical decision support systems. In: Collection of technical papers—AIAA 1st intelligent systems technical conference, vol 2, pp 557–562. https://doi.org/10.2514/6.2004-6313

Dikmen M, Burns C (2022) The effects of domain knowledge on trust in explainable AI and task performance: a case of peer-to-peer lending. Int J Hum Comput Stud 162:102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Dellermann D, Calma A, Lipusch N, Weber T, Weigel S, Ebel P (2021) The future of human–AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. arXiv:2105.03354

Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mähringer-Kunz A, Püsken M, Baeßler B, Sauer S, Maintz D, Pinto Dos Santos D (2023) Automation bias in mammography: the impact of artificial intelligence BI-RADS suggestions on reader performance. Radiology 307(4):e222176. https://doi.org/10.1148/radiol.222176

Duan W, Zhou S, Scalia MJ, Yin X, Weng N, Zhang R, Freeman G, McNeese N, Gorman J, Tolston M (2024) Understanding the evolvement of trust over time within human–AI teams. Proc ACM Hum Comput Interact 8(CSCW2):1–31

Gaube S, Suresh H, Raue M, Merritt A, Berkowitz SJ, Lermer E, Coughlin JF, Guttag JV, Colak E, Ghassemi M (2021) Do as AI say: susceptibility in deployment of clinical decision-aids. NPJ Dig Med 4(1):31. https://doi.org/10.1038/s41746-021-00385-9

Ge X, Xu C, Misaki D, Markus HR, Tsai JL (2024) How culture shapes what people want from AI. In: Proceedings of the 2024 CHI conference on human factors in computing systems, vol 95, pp 1–15. https://doi.org/10.1145/3613904.3642660

Gegoff I, Tatasciore M, Bowden VK, Loft S (2025) Deciphering automation transparency: do the benefits of transparency differ based on whether decision recommendations are provided? Hum Fact. https://doi.org/10.1177/00187208251318465

Glick A, Clayton M, Angelov N, Chang J (2022) Impact of explainable artificial intelligence assistance on clinical decision-making of novice dental clinicians. JAMIA Open 5(2):ooac031. https://doi.org/10.1093/jamiaopen/ooac031

Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089

Hemmer P, Schemmer M, Vössing M, Kühl N (2021) Human–AI complementarity in hybrid intelligence systems: a structured literature review. In: Pacific Asia conference on information systems, vol 78. ISBN 978-1-7336325-7-7

Horowitz MC, Kahn L (2024) Bending the automation bias curve: a study of human and AI-based decision making in national security contexts. Int Stud Q 68(2):020. https://doi.org/10.1093/isq/sqae020

Jacobs M, Pradier MF, McCoy TH Jr, Perlis RH, Doshi-Velez F, Gajos KZ (2021) How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Transl Psychiatry 11(1):108. https://doi.org/10.1038/s41398-021-01224-x

Jones-Jang SM, Park YJ (2023) How do people react to AI failure? Automation bias, algorithmic aversion, and perceived controllability. J Comput Mediat Commun 28(1):zmac029. https://doi.org/10.1093/jcmc/zmac029

Keding C, Meissner P (2021) Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. Technol Forecast Soc Change 171:120970. https://doi.org/10.1016/j.techfore.2021.120970

Kim SH, Schramm S, Riedel EO, Schmitzer L, Rosenkranz E, Kertels O, Bodden J, Paprottka K, Sepp D, Renz M, Kirschke J, Baum T, Maegerlein C, Boeckh-Behrens T, Zimmer C, Wiestler B, Hedderich DM (2025) Automation bias in AI-assisted detection of cerebral aneurysms on time-of-flight MR angiography. Radiol Med (Torino) 130(4):555–566. https://doi.org/10.1007/s11547-025-01964-6

Klingbeil A, Grützner C, Schreck P (2024) Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. Comput Hum Behav 160:108352

Kücking F, Hübner U, Przysucha M, Hannemann N, Kutza JO, Moelleken M, Busch D (2024) Automation bias in AI-decision support: Results from an empirical study. Ger Med Data Sci (pp. 298–304). IOS Press

Küper A, Lodde GC, Livingstone E, Schadendorf D, Krämer N (2025) Psychological factors influencing appropriate reliance on ai-enabled clinical decision support systems: experimental web-based study among dermatologists. J Med Internet Res 27:e58660. https://doi.org/10.2196/58660

Kupfer C, Prassl R, Fleiß J, Malin C, Thalmann S, Kubicek B (2023) Check the box! How to deal with automation bias in AI-based personnel selection. Front Psychol 14:1118723. https://doi.org/10.3389/fpsyg.2023.1118723

Lai V, Chen C, Liao QL, Smith-Renner A, Tan C (2023) Towards a science of human–AI decision making: an overview of Design space in empirical human-subject studies. In: 2023 ACM conference on fairness, accountability, and transparency (FAccT '23). https://doi.org/10.48550/arXiv.2112.11471

Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Fact 46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Long D, Magerko B (2020) What is AI literacy? Competencies and design considerations. In: Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, pp 1–16. https://doi.org/10.1145/3313831.3376727

Lu Z, Yin M (2021) Human reliance on machine learning models when performance feedback is limited: heuristics and risks. In: Proceedings of the 2021 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, vol 78, pp 1–16. https://doi.org/10.1145/3411764.3445562

Lyell D, Coiera E (2017) Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 24(2):423–431. https://doi.org/10.1093/jamia/ocw105

Matzen LE, Gastelum ZN, Howell BC, Divis KM, Stites MC (2024) Effects of machine learning errors on human decision-making: manipulations of model accuracy, error types, and error

importance. Cogn Res Princ Implic 9(1):56. https://doi.org/10.1186/s41235-024-00586-2
Naiseh M, Al-Thani D, Jiang N, Ali R (2023) How the different explanation classes impact trust calibration: The case of clinical decision support systems. Int J Hum Comput Stud 169:102941. https://doi.org/10.1016/j.ijhcs.2022.102941
Nourani M, Roy C, Block JE, Honeycutt DR, Rahman T, Ragan E, Gogate V (2022) On the importance of user backgrounds and impressions: lessons learned from interactive AI applications. ACM Trans Interact Intell Syst 12(4):28. https://doi.org/10.1145/3531066. (**1–29**)
Okamura K, Yamada S (2020) Adaptive trust calibration for human–AI collaboration. PLoS ONE 15(2):e0229132. https://doi.org/10.1371/journal.pone.0229132
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ (Clin Res Ed). https://doi.org/10.1136/bmj.n71
Parasuraman R, Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. Hum Fact 52(3):381–410. https://doi.org/10.1177/0018720810376055
Rastogi C, Zhang Y, Wei D, Varshney KR, Dhurandhar A, Tomsett R (2022) Deciding fast and slow: the role of cognitive biases in ai-assisted decision-making. Proc ACM Hum Comput Interact 6(CSCW1):1–22
Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB (2021) PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev 10:1–19. https://doi.org/10.1186/s13643-020-01542-z
Rezazade Mehrizi MH, Mol F, Peter M, Ranschaert E, Dos Santos DP, 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. Sci Rep 13(1):9230. https://doi.org/10.1038/s41598-023-36435-3
Romeo G, Conti D (2024) Beyond automation: reshaping Human-artificial intelligence interaction. Sistemi Intelligenti XXVI(3):641–648. https://doi.org/10.1422/115336
Ruschemeier H, Hondrich LJ (2024) Automation bias in public administration—an interdisciplinary perspective from law and psychology. Gov Inf Q 41(3):101953. https://doi.org/10.1016/j.giq.2024.101953
Sato T, Yamani Y, Liechty M, Chancey ET (2020) Automation trust increases under high-workload multitasking scenarios involving risk. Cogn Technol Work 22:399–407. https://doi.org/10.1007/s10111-019-00580-5
Selten F, Robeer M, Grimmelikhuijsen S (2023) "Just like I Thought": street-level bureaucrats trust AI recommendations if they confirm their professional judgment. Public Admin Rev 83(2):263–278. https://doi.org/10.1111/puar.13602
Stern C, Jordan Z, McArthur A (2014) Developing the review question and inclusion criteria. Am J Nurs 114(4):53–56. https://doi.org/10.1097/01.NAJ.0000445689.67800.86
Steyvers M, Kumar A (2024) Three challenges for AI-assisted decision-making. Perspect Psychol Sci 19(5):722–734. https://doi.org/10.1177/17456916231181102
Tatasciore M, Loft S (2025) Calibrating reliance on automated advice: transparency and trust calibration feedback. Int J Hum Comput Interact. https://doi.org/10.1080/10447318.2025.2487861
Vasconcelos H, Jörke M, Grunde-McLaughlin M, Gerstenberg T, Bernstein MS, Krishna R (2023) Explanations can reduce overreliance on AI systems during decision-making. In: Proceedings of the ACM on human-computer interaction, 7 (CSCW1), vol 129, pp 1–38. https://doi.org/10.1145/3579605
Vered M, Livni T, Howe PDL, Miller T, Sonenberg L (2023) The effects of explanations on automation bias. Artif Intell 322:103952. https://doi.org/10.1016/j.artint.2023.103952
Wickens CD, Clegg BA, Vieane AZ, Sebok AL (2015) Complacency and automation bias in the use of imperfect automation. Hum Fact 57(5):728–739. https://doi.org/10.1177/0018720815581940
Wickens CD, Helton WS, Hollands JG, Banbury S (2021) Chapter 13 | human–automation interaction. In: Engineering psychology and human performance, 5th edn. Routledge, pp 516–551. https://doi.org/10.4324/9781003177616
Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (FAT* '20). Association for Computing Machinery, New York, NY, USA, pp 295–305. https://doi.org/10.1145/3351095.3372852
Zhang ZT, Argın SK, Bilen MB, Urgun D, Deniz SM, Liu Y, Hassib M (2024) Measuring the effect of mental workload and explanations on appropriate AI reliance using EEG. Behav Inf Technol. https://doi.org/10.1080/0144929X.2024.2431055