Calibrating Reliance: Addressing Misuse and Disuse in AI-Based Second-Opinion Systems for Medical Diagnosis

Federico Cabitza[1,2], **Andrea Campagner**[1,2], **Gian Eugenio Tontini**[3,4],
[1]University of Milano-Bicocca, Viale Sarca 336, 20126, Milan, Italy
[2]IRCCS Ospedale Galeazzi-Sant'Ambrogio, Via Cristina Belgioioso, 173, 20157, Milan, Italy
[3]University of Milan, Via Festa del Perdono, 7, 20122 Milan, Italy
[4]Policlinico Hospital of Milan, Via Francesco Sforza, 35, 20122, Milan, Italy
federico.cabitza@unimib.it, andrea.campagner@unimib.it, gianeugeniotontini@gmail.com

Abstract

AI systems are widely proposed as second-opinion advisors in clinical diagnosis, offering the promise of enhancing decision accuracy and clinician confidence while preserving human oversight. However, successful deployment in real-world practice faces a critical barrier: clinicians' reliance on AI is often miscalibrated, manifesting as misuse (over-reliance driven by automation bias) and disuse (underutilization driven by self-anchoring bias). This paper addresses these deployment challenges by systematically analyzing how such reliance patterns affect diagnostic accuracy, confidence, and decision-making across diverse medical specialties. We report results from controlled simulations involving over 300 medical professionals across six diagnostic settings—including knee MRI analysis, spinal X-rays, cardiac ECG evaluation, and gastrointestinal endoscopy—using a human-first, AI-second workflow. Although AI advice improved average diagnostic accuracy (+2 percentage points) and clinician confidence (+3 points on a normalized scale), overall levels of appropriate reliance remained well below 50%, with disuse emerging as the more prevalent and consequential barrier. We introduce and validate Appropriate Reliance as an actionable metric for assessing and improving human-AI collaboration, providing practical guidance for developers, healthcare institutions, and policymakers seeking to deploy second-opinion AI systems safely and effectively. By identifying the sociotechnical barriers and offering evidence-based design insights, this work supports the emerging application of AI as a collaborative advisor in clinical workflows, charting a clear path toward deployment that enhances diagnostic safety, accountability, and patient care. Specifically, we propose integrating the Appropriate Reliance metric into system development workflows, clinician training, and regulatory evaluations to enable safe and effective deployment of second-opinion AI systems.

Software —
https://www.entechne.com/metimeter/haiassessment

Introduction

This contribution addresses an *emerging application* of AI in medicine: the use of computational systems as second-opinion advisors in complex diagnostic contexts. While fully autonomous AI diagnosis remains ethically problematic and is subject to stringent regulatory constraints (Iniesta 2023), AI-based second opinions are widely regarded as a promising and feasible mode of deployment. Such systems can augment clinical decision-making while preserving human oversight, aligning with regulatory requirements and ethical expectations, and offering a feasible path for near-term deployment in routine practice (Buçinca, Malaya, and Gajos 2021; Brodeur et al. 2024).

However, deploying these systems in real-world clinical practice faces persistent challenges. Notably, clinicians' reliance on AI is often *miscalibrated*, manifesting as either over-reliance (*misuse*) or under-utilization (*disuse*). These forms of inappropriate use can undermine the potential benefits of AI integration, complicate accountability, and limit adoption (Parasuraman and Riley 1997).

Over the past four years (2021-2025), we have conducted a series of simulation-based studies to systematically investigate these barriers to effective deployment. In controlled experiments involving more than 300 medical professionals across diverse diagnostic specialties, participants engaged with AI systems of varying accuracy to replicate realistic second-opinion workflows. These studies generated a rich dataset capturing decision-making accuracy, confidence, and nuanced patterns of human-AI interaction.

While some results from these comparative studies have been published previously (Cabitza et al. 2022, 2023b,a), this paper takes a new step by offering a consolidated, cross-study analysis focused specifically on quantifying misuse and disuse in medical second-opinion contexts. We also introduce and validate *Appropriate Reliance* as an engineering-oriented metric that can be integrated into system design, interface evaluation, clinician training, and regulatory assessment to guide deployment of AI systems.

By addressing these questions, this work aims to reduce the gap between promising laboratory demonstrations and sustainable, safe clinical adoption. We also outline concrete next steps for deployment, including validating the Appropriate Reliance metric in real-world settings and embedding it into iterative design cycles for AI decision support tools. Our findings provide evidence-based insights for designing AI advisory systems that promote calibrated trust and maximize clinical benefit—essential prerequisites for the successful deployment of AI systems in routine healthcare.

Methods

To operationalize the concepts of *misuse* (overreliance on automation) and *disuse* (underutilization of automation), it is necessary to define corresponding appropriate metrics. For this purpose, we adopt the framework proposed in (Cabitza et al. 2023b), which conceptualizes second opinion settings as consisting of three interpretive or decision-making moments (see Figure 1): the initial human decision (HD1), the advice or opinion provided by the machine (AI), and the final decision made by the human after considering the machine's suggestion (FHD).

These three decision-making interactions define eight distinct conditions, hereafter referred to as *reliance patterns*, determined by whether each individual interpretation is correct (1) or not (0). We represent these patterns as the set of all ordered 3-tuples over {1,0} (i.e., $000, 001, 010, 011, 100, 101, 110, 111$), where each digit indicates the correctness of one interpretation. Identifying and classifying reliance patterns across instances of human–machine interaction enables systematic quantification of phenomena typical of second-opinion settings. For example, one can count the frequency with which each pattern occurs (denoted, e.g., N_{001} for pattern 001); the total number of such interactions is denoted as N_{***}.

That said, we define *appropriate reliance* (AR) as the proportion of cases in which humans trust the machine and follow its advice when it is correct, and distrust the machine and disregard its advice when this is incorrect. Referring to the reliance patterns outlined above, *appropriate reliance* is calculated as the rate of successful acts of reliance:[1])

$$AR = \frac{N_{011} + N_{101} + N_{001} + N_{000}\downarrow + N_{111}\uparrow}{N_{***}} \quad (1)$$

Likewise, *deference* is defined as the proportion of instances in which humans revise their decisions to align with the machine's opposing advice, expressed as $(N_{100} + N_{011})/N_{XY*}$. Conversely, *determination* denotes the proportion of cases in which humans retain their initial decision despite the machine's contradictory advice, expressed as $(N_{101} + N_{010})//N_{XY*}$.

These two latter cases can further be categorized based on whether the final decision was correct or incorrect (i.e., $**1$ or $**0$). More in particular, *misuse* occurs when *deference* results in an error that would not have occurred had the human decision makers adhered to their initial opinion and ignored the machine's advice (that is, N_{100} corresponds to the number of misuse cases). Conversely, *disuse* occurs when *determination* leads to an error that could have been avoided by following the machine's advice (that is, N_{010} corresponds to number of disuse cases).

Misuse is often linked to *automation bias*, the tendency to defer to machine-generated advice, which has been extensively studied in the literature. In this context, we operationalize *automation bias* using Formula 2.a, which calculates the log-odds ratio between the 100 and 101 patterns. A higher value of this log-odds indicates a greater degree of *automation bias*.

$$\begin{aligned} &\text{a)} \quad AB = \log\left(\frac{\frac{N_{100}}{N_{***}-N_{100}}}{\frac{N_{101}}{N_{***}-N_{101}}}\right) \\ &\text{b)} \quad SAB = \log\left(\frac{\frac{N_{010}}{N_{***}-N_{010}}}{\frac{N_{011}}{N_{***}-N_{011}}}\right) \end{aligned} \quad (2)$$

Disuse, on the other hand, arises from an opposite bias, similar to systematic tendencies observed in Judge-Advisor Systems (Harvey and Fischer 1997). In such cases, decision makers (acting as judges) fail to integrate the input of external advisors into their final decisions. While this bias is less frequently studied in human-computer interaction contexts, it has been referred to as *conservatism bias* in (Cabitza et al. 2023a) or *fixation* in (Klein 2022). In the cognitive sciences literature, this phenomenon has also been described using various terms, including *egocentric advice discounting* (Bonaccio and Dalal 2006), *self/other effect* (Yaniv 2004), *cognitive rigidity* (Schmitt et al. 2021), *cognitive inertia* (Alós-Ferrer, Hügelschäfer, and Li 2016), *belief perseverance* (Soper 2020), and *self-generated anchoring effect* (Li et al. 2010; Fogliato et al. 2022). Here, we refer to this phenomenon as *self-anchoring bias* and operationalize it using Formula 2.b, which calculates the log-odds ratio between the rates the 010 and 011 patterns occur. A higher value of this log-odds indicates a greater degree of *self-anchoring bias*. The variances of the log-odds (in the form $log((a/b)/(c/d))$, and equivalently $log(ad/bc)$) were calculated using the approximation $1/a + 1/b + 1/c + 1/d$.

In what follows, we outline the main characteristics of the studies in which patterns of misuse and disuse were observed in simulations of medical second-opinion settings. The common element in all studies is the adoption of the second-opinion protocol, whereby participants first recorded their initial diagnosis (HD1), then reviewed the AI's recommendation before making their final determination (FHD). Diagnostic confidence was assessed using a 6-point ordinal scale (1: no confidence; 6: complete confidence) for both the initial (C1) and final (FC) decisions and collected along the case interpretations HD1 and FHD through a multi-page

[1]We adopt the metrics proposed in (Cabitza et al. 2025), which also considers the patterns 000 and 111 by counting the number of times the former one is associated with a decrease in confidence (↓) and the latter one with an increase of confidence (↑).

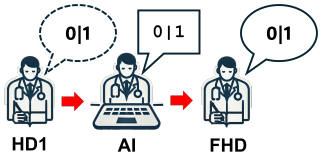

Figure 1: Protocol for collecting reliance patterns, where HD1 denotes the initial human decision or interpretation, AI represents the machine's advice, and FHD signifies the final human decision after consulting the AI's second opinion. Each of these elements can independently be either correct (1) or incorrect (0). The patterns of these temporal sequence can be represented by eight ordered 3-tuples over {0,1}.

LimeSurvey questionnaire.

In the Knee MRI Analysis Study (MRI) study, a cohort of 12 board-certified radiologists was recruited from IRCCS Ospedale Galeazzi Sant'Ambrogio in Milan and other Italian medical centers to participate in AI-assisted knee lesion classification. The investigation utilized 120 cases from the MRNet dataset selected by a board-certified radiologist for their representativeness of challenging cases. The AI was simulated to be 80% accurate. Additional methodological details are provided in Cabitza et al. (2023).

The Spinal X-Ray Diagnosis Study (XRAYS) investigated traumatic thoraco-lumbar fracture detection with seven orthopedic specialists of varying expertise levels called to evaluate 12 X-ray images. The AI was simulated to be 78% accurate. Further procedural details are documented in Cabitza et al. (2022).

In the Cardiac ECG Evaluation Study (ECG) study, twenty-one medical professionals affiliated with the University Hospital of Siena's Medical School in Italy participated in a study analyzing cardiac rhythm patterns. The research protocol involved the classification of 20 ECG recordings, carefully selected by a cardiologist from the ECG Wave-Maven repository. Complete methodological details are provided in Cabitza et al. (2023).

The Gastrointestinal Endoscopy Assessment Study (ENDO) study is a large-scale endoscopic evaluation that engaged 274 medical professionals all over Europe across three studies: ENDO-HP, ENDO-IL, and ENDO-UCEIS. In the ENDO-HP study, participants assessed 30 cases of potential gastrointestinal bleeding (Hemorrhagic Potential) in short (15–20 second) endoscopic colonoscopy videos. In ENDO-IL, participants evaluated 30 cases of potential inflammatory lesions in small bowel capsule endoscopy. In ENDO-UCEIS, participants assessed and graded 30 cases of ulcerative colitis activity using proctosigmoidoscopy. In all studies the AI was 80% accurate. Methodological details are provided in Tontini et al. (under review).

In the next section, we will present the findings from the studies described above, with a focus on the phenomena of misuse and disuse. The AR, AB and SAB scores have been computed with an online tool that we had developed to assess several dimensions of human-AI interaction (Natali, Campagner, and Cabitza 2024). These findings will be analyzed both at the level of individual studies (as reported in the cited articles) and at an aggregated level, using variance- and sample size-weighted analysis, as is customary in meta-analytic approaches (Borenstein et al. 2021).

To quantitatively synthesize the results across studies, we performed a meta-analysis for each metric, using either the observed rate (e.g., accuracy, appropriate reliance) or the log odds ratio, as appropriate. For each study, the point estimate and its variance were calculated under the assumption of a within-subjects design. In order to account for the dependence among repeated measurements within the same rater or subject, the variance of each proportion was corrected by the intra-rater correlation coefficient, set at 0.75 in accordance with the structure of the data. Specifically, the variance for each proportion p was estimated as $\mathrm{Var}(p) = p(1-p)/[N(1-\rho)]$, where N is the number of subjects and ρ is the intra-rater correlation.

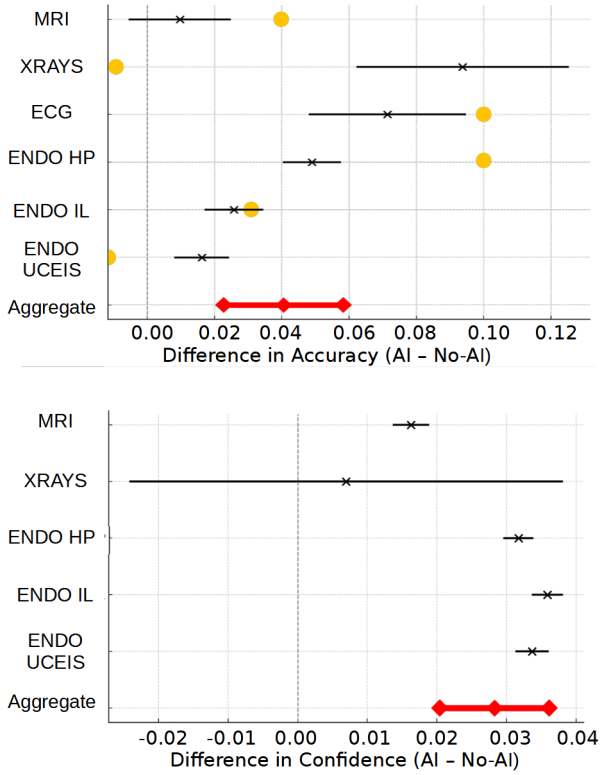

Figure 2: Forest plots summarizing user study results. The top panel shows differences in diagnostic accuracy between post-AI and baseline (pre-AI) conditions for each study and the random-effects aggregate. Orange circles represent the accuracy difference between the simulated AI and the average baseline human reader (positive values favor AI). The bottom panel shows corresponding differences in diagnostic confidence.

For each metric, we pooled the study-specific estimates using a random-effects meta-analytic model (DerSimonian-Laird method), which incorporates both within-study and between-study variance. The choice between a fixed-effect and a random-effects model was informed by the observed heterogeneity, as quantified by the I^2 statistic. When substantial heterogeneity was present ($I^2 > 50\%$), the random-effects model was preferred, as it provides a more conservative and generalizable estimate by allowing for genuine differences in effect size across studies. The pooled effect and its 95% confidence interval are reported for each metric. In the case of rate-based metrics, the analysis was conducted directly on the proportions, while for odds ratios, the analysis was performed on the logarithm scale to stabilize variance and normalize the sampling distribution.

This approach ensures that the meta-analytic estimates ac-

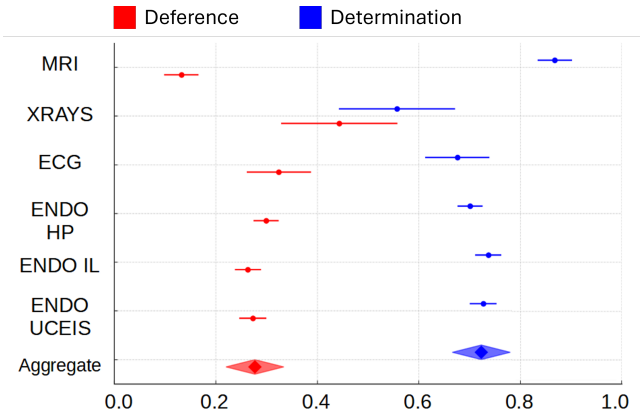

Figure 3: Forest plot illustrating deference (red, bottom) to AI advice and determination (blue, top) in affirming the initial decision.

curately reflect both the within-study precision and the heterogeneity observed in the underlying evidence, and that the statistical inference remains valid despite the correlated nature of the within-subject data.

Results

The results of the meta-analyses are represented in tabular and visual format, respectively in Table 1, Figure 2, Figure 3, Figure 4 and Figure 5.

In particular, the meta-analyses evaluating the effect of AI assistance on diagnostic accuracy and confidence are summarized in Figure 2. For diagnostic accuracy (six studies), the difference (AI minus No-AI) was computed using a within-subject design with intra-subject correlation set at 0.75. Observed accuracy differences ranged from 0.0097 (MRI) to 0.0937 (XRAYS), with significant improvements in four studies. Between-study heterogeneity was substantial ($I^2 = 90.1\%$); thus, a random-effects model was used, yielding a statistically significant pooled effect (mean difference = 0.035, 95% CI: 0.019–0.051).

For diagnostic confidence (five studies), the mean difference was estimated using an intra-rater correlation of 0.89. Observed differences ranged from 0.0069 (XRAYS) to 0.0359 (ENDO-IL), with significant positive effects in four out of five studies. Heterogeneity was high ($I^2 = 97.6\%$), and the pooled estimate indicated a small but significant increase (mean difference = 0.0283, 95% CI: 0.0204–0.0361).

In both analyses, forest plots display individual study effects and confidence intervals in order, with the overall pooled effect as a diamond at the bottom.

Meta-analyses of automation bias and self-anchoring bias, shown in Figure 5, used log-odds ratios with standard variance approximations. Automation bias showed consistently negative log-odds ratios, with a significant pooled reduction (–1.83, 95% CI: –2.15 to –1.51; $I^2 = 79.3\%$). Self-anchoring bias showed a consistently positive effect (pooled log-odds ratio = 0.56, 95% CI: 0.18–0.95; $I^2 = 94.0\%$). Again, forest plots display study-level and aggregate effects.

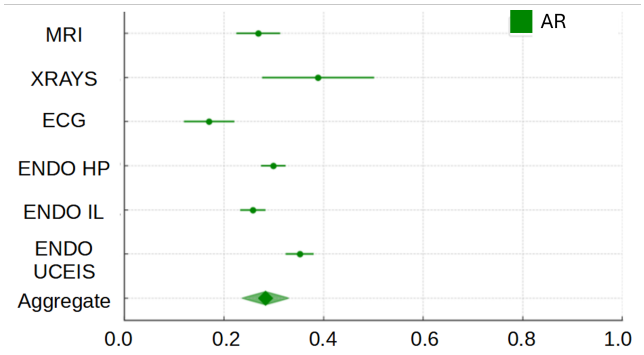

Figure 4: Forest plot illustrating the appropriate reliance scores, across all individual studies as well as at the pooled level.

For deference and determination (Figure 3), proportions and 95% CIs were calculated for each study, adjusting variance for intra-rater correlation (0.75). Random-effects pooling (DerSimonian-Laird method) accounted for substantial heterogeneity ($I^2 = 93.7\%$). The pooled estimates were 0.28 (95% CI: 0.22–0.33) for deference and 0.72 (95% CI: 0.67–0.78) for determination, indicating a general tendency toward determination over deference. Forest plots show both metrics for each study and the pooled effects as diamonds.

The meta-analysis of appropriate reliance, also in Figure 3, was performed analogously. Proportions were calculated with variance adjusted for intra-rater correlation, and pooled using a random-effects model ($I^2 = 90.0\%$). The pooled estimate was 0.28 (95% CI: 0.24–0.33), reflecting high between-study variability.

Overall, forest plots for all outcomes follow a consistent format: individual study estimates and confidence intervals are shown in order, with pooled effects presented as diamonds at the bottom of each plot. Across outcomes, substantial heterogeneity highlights the variability in effect sizes, but consistent directions of effect for AI assistance and measured biases are observed.

Discussion

Medicine is among the professional domains that have most fully embraced simulation as a critical tool for workforce training (Barry Issenberg et al. 2005). Simulation offers exercises that closely replicate real-world cases, requiring skillful and timely decision-making. This approach serves two key purposes. First, it provides practitioners with a controlled environment where errors have minimal consequences, enabling analysis of failures and the development of strategies to mitigate similar risks in real-world settings. Second, it allows for a deeper exploration of phenomena before they produce tangible effects on human health or well-being (Cook and Triola 2009). Our study focuses on this second function: using simulation–specifically involving real practitioners working on real cases–as a tool to investigate phenomena related to the impact of AI systems in decision-making contexts, thereby supporting more informed and responsible deployment of these systems in such settings.

Study	Deference	Determination	Automation Bias	Self-Anchoring Bias	Appropriate Reliance
MRI	0.13 [0.12, 0.15]	0.87 [0.85, 0.88]	-2.10 [-3.10, -1.10]	1.80 [1.20, 2.40]	0.27 [0.24, 0.30]
XRAYS	0.44 [0.39, 0.50]	0.56 [0.50, 0.61]	-2.00 [-3.00, -1.00]	-1.30 [-2.30, -0.30]	0.39 [0.33, 0.45]
ECG	0.32 [0.29, 0.36]	0.68 [0.64, 0.71]	-1.50 [-2.60, -0.40]	0.20 [-0.50, 0.90]	0.17 [0.15, 0.19]
ENDO-HP	0.30 [0.29, 0.31]	0.70 [0.69, 0.71]	-1.60 [-2.10, -1.10]	0.40 [-0.20, 1.00]	0.30 [0.29, 0.31]
ENDO-IL	0.26 [0.25, 0.28]	0.74 [0.72, 0.75]	-1.70 [-2.20, -1.20]	0.70 [0.10, 1.30]	0.26 [0.24, 0.27]
ENDO-UCEIS	0.27 [0.26, 0.29]	0.73 [0.71, 0.74]	-1.20 [-1.70, -0.70]	0.80 [0.20, 1.40]	0.35 [0.34, 0.37]
Pooled	**0.28 [0.22, 0.33]**	**0.72 [0.67, 0.78]**	**-1.50 [-1.70, -1.30]**	**0.70 [0.50, 0.90]**	**0.28 [0.24, 0.33]**

Table 1: Meta-analytic summary of deference, determination, automation bias, self-anchoring bias, and appropriate reliance across studies. All values are shown as point estimate [95% CI]. The bottom row shows the pooled effect for each metric.

In our pre-deployment simulation, we considered it important to look beyond accuracy differences, which are typically the focus of post-deployment studies. Our analysis consistently showed an increase in accuracy, although in one study (MRI) this increase was not statistically significant. Notably, this also occurred when the AI was less accurate than the average rater, as in the XRAYS and ENDO-UCEIS studies. These are examples of the so-called *spur* effect (Cabitza 2025), which occurs when users improve their accuracy despite the AI being less accurate–a sign of support effectiveness that is decoupled from mere accuracy.

Moreover, our analysis highlighted another important aspect related to the third pillar of usability (in addition to effectiveness and efficiency), i.e., satisfaction. A proxy for this often-neglected dimension is users' confidence in their own decisions. In this case, the decision support—while it contradicted the initial human opinion in 27% of cases—had a significant pooled effect in increasing users' self-confidence.

However, we also argue that pre-deployment analysis should consider other dimensions beyond accuracy, such as the extent to which opportunistic behaviors (e.g., deference) or advice rejection (i.e., determination) lead to dysfunctional outcomes. We operationalize these phenomena through metrics quantifying automation bias and self-anchoring bias. Evaluations using these metrics can be complemented by measures of appropriate reliance, which offer a more pragmatically useful indication: whether users rely on AI as a second-opinion advisor appropriately. In this regard, we provide the qualitative classification depicted in 2, where values below 30% may be interpreted as poor agreement and values below 20% as indicating the need for corrective measures. This includes considering what information the system provides to users (e.g., advice with or without uncertainty indications such as confidence scores, with or without explanations), what vendors disclose about the system's capabilities (shaping its reputation and transparency), and, last but not least, the necessary training for users to help them either learn to trust more—if self-anchoring bias is the problem—or evaluate information more critically if the opposite problem, automation bias, is more severe.

The combined effect is poor, indicating clear room for improvement through a combination of interventions such as those suggested above. This result stems from studies where appropriateness was at the margins of adequacy (i.e., XRAYS) and studies that were nearly significantly below the threshold of acceptability (i.e., ECG). Whenever a pre-deployment simulation or pilot study yields a low score, developers and deployers should collaborate to investigate the main causal factors, including the analysis of deference, determination, and the biases that foster misuse and disuse.

Appropriate Reliability score	Classification
$AR \geq 0.6$	Excellent
$0.5 \leq AR < 0.6$	Very Good
$0.4 \leq AR < 0.5$	Good
$0.3 \leq AR < 0.4$	Fair
$0.2 \leq AR < 0.3$	Poor
$AR < 0.2$	Insufficient

Table 2: Interpretation classification for Appropriate Reliance levels.

In particular, across all studies we conducted, automation bias was negligible (see Figure 5), and thus misuse was minimal, as the log-odds were all significantly below the nil effect line, in fact even lower than -1. This means the odds of exhibiting automation bias were less than 37% of the odds of confirming the correct decision when exposed to incorrect advice. Conversely, disuse was consistently observed in all studies except XRAYS, with especially notable effects in the MRI study. The pooled effect is 0.7, indicating that self-anchoring bias is associated with roughly twice the odds of failing to change an initial incorrect decision after being exposed to correct advice. As shown in Figure 3, appropriate reliance is higher when deference and determination are balanced and both biases are absent.

Although the meta-analyses revealed substantial heterogeneity across studies (I^2 values above 80%), this variability primarily reflects the diversity of diagnostic tasks, participant backgrounds, and AI configurations examined, rather than inconsistencies in methodological quality. While such heterogeneity may limit the strict comparability of individual results and reduce the precision of pooled estimates, it also strengthens the ecological validity of the findings by encompassing a wide range of realistic clinical contexts.

Path Toward Deployment

In this subsection, we generalize the assessment procedure illustrated above, based on the user studies we conducted in six different diagnostic settings. To facilitate real-world

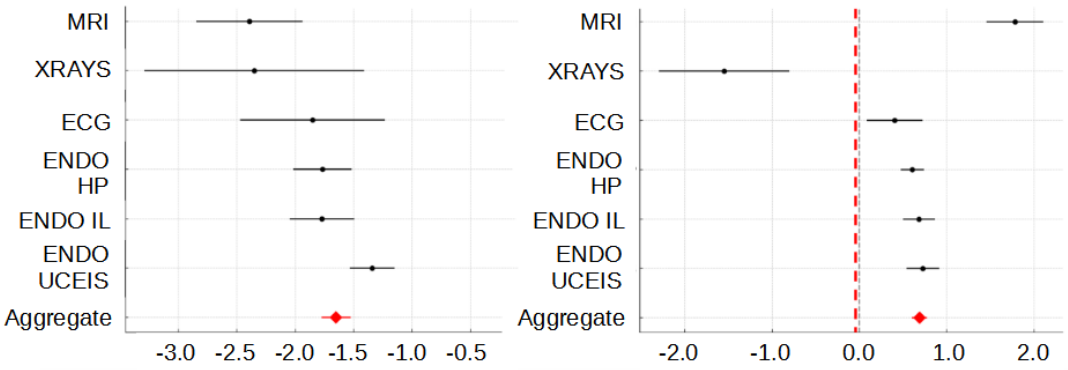

Figure 5: Forest plots illustrating automation bias (left), self-anchoring bias (right), across all individual studies.

adoption, we propose the following structured roadmap for deployment. First, integrating the Appropriate Reliance (AR) metric into iterative design processes for AI decision support systems will enable developers to systematically evaluate and refine user interfaces, feedback mechanisms, and workflow integrations that promote calibrated trust. As shown above, AR must be assessed together with AB and SAB, as these are the main factors contributing to low AR scores. Second, validating the predictive utility of the AR metric in live clinical environments is essential to ensure its generalizability beyond controlled simulations, enabling robust assessment of its impact on diagnostic accuracy, decision-making consistency, and clinician confidence across diverse specialties and healthcare settings. Third, developing targeted training interventions informed by reliance will directly address the sociotechnical barriers of misuse and disuse, equipping clinicians with the skills to recognize, interpret, and appropriately integrate AI advice. Such interventions should include scenario-based simulations, decision-making exercises, and explanatory modules on AI strengths and limitations. Fourth, close collaboration with regulatory bodies and professional organizations will be critical to establish reliance calibration as a recognized criterion for the approval and certification of second-opinion AI systems, ensuring alignment with safety, accountability, and ethical standards. Together, these steps offer a concrete, evidence-based pathway to translate research findings into deployable solutions that are not only technically effective but also socially and clinically sustainable.

Conclusion

This work set out to address a critical barrier to the successful deployment of AI-based second-opinion systems in clinical diagnosis: the miscalibration of clinician reliance, manifesting as misuse (over-reliance from automation bias) and disuse (under-utilization from self-anchoring bias). Through a meta-analytic synthesis of large-scale simulation studies across diverse diagnostic contexts, we quantified these biases and introduced Appropriate Reliance (AR) as a practical, actionable metric to guide the design and evaluation of human-AI collaboration in healthcare. Disuse—marked by obstinance and failure to adopt correct AI advice—emerges as a more prevalent and consequential barrier than automation bias. This challenges the prevailing focus on over-reliance as the main safety concern in AI-assisted decision-making and highlights the need for a broader, evidence-based approach to system design and training.

These insights are not merely theoretical. They offer concrete engineering guidance for developers, providers, and policymakers on how to design, evaluate, and deploy second-opinion AI systems in clinical practice. Promoting calibrated trust will require more than improving model accuracy or interface transparency. It will demand investments in explainability, robust validation, clear communication of uncertainty, and fostering user accountability—essential for aligning human and machine reasoning in high-stakes settings. Our AR metric provides a systematic way to assess these goals and track progress toward safe, effective integration. While our studies used controlled simulations designed to approximate real-world workflows, validating these insights in live clinical environments remains a crucial next step. Future research should focus on deploying these metrics in production systems, evaluating interventions to reduce self-anchoring bias, and tailoring solutions for diverse clinician groups. Future work will also include a systematic comparison between the Appropriate Reliance metric and existing measures of usability, trust, and reliance in human–AI interaction—such as those derived from the ISO 9241-210 framework—to clarify their respective scopes, overlaps, and complementarity in clinical evaluation.

Ultimately, this work supports the emerging application of AI as a collaborative advisor in medicine—a role that complements human expertise rather than replacing it. Future work will integrate the AR metric into production-ready systems, conduct in-situ clinical evaluations, and refine user training to ensure safe, effective, and accountable deployment. By systematically improving the conditions for appropriate reliance, we can help ensure that AI deployment in healthcare delivers on its promise of more accurate, accountable, and equitable diagnostic decision-making that enhances patient care.

References

Alós-Ferrer, C.; Hügelschäfer, S.; and Li, J. 2016. Inertia and decision making. *Frontiers in psychology*, 7: 169.

Barry Issenberg, S.; Mcgaghie, W. C.; Petrusa, E. R.; Lee Gordon, D.; and Scalese, R. J. 2005. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review. *Medical teacher*, 27(1): 10–28.

Bonaccio, S.; and Dalal, R. S. 2006. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational behavior and human decision processes*, 101(2): 127–151.

Borenstein, M.; Hedges, L. V.; Higgins, J. P.; and Rothstein, H. R. 2021. *Introduction to meta-analysis*. John Wiley & Sons.

Brodeur, P. G.; Buckley, T. A.; Kanjee, Z.; Goh, E.; Ling, E. B.; Jain, P.; Cabral, S.; Abdulnour, R.-E.; Haimovich, A.; Freed, J. A.; et al. 2024. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*.

Buçinca, Z.; Malaya, M. B.; and Gajos, K. Z. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1): 1–21.

Cabitza, F. 2025. Who Knocks on Heaven's Door: Measuring Augmentation and Outperformance in Human–AI Diagnostic Teams. In *International Conference on Artificial Intelligence in Medicine*, 73–77. Springer.

Cabitza, F.; Campagner, A.; Angius, R.; Natali, C.; and Reverberi, C. 2023a. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, 1–20.

Cabitza, F.; Campagner, A.; Famiglini, L.; Gallazzi, E.; and La Maida, G. A. 2022. Color shadows (part i): Exploratory usability evaluation of activation maps in radiological machine learning. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, 31–50. Springer.

Cabitza, F.; Campagner, A.; Fregosi, C.; Cameli, M.; Gallazzi, E.; Sconfienza, L. M.; and Tontini, G. E. 2025. Five Degrees of Separation: Investigating the Unexpected Potential of Displaced Human-AI Collaboration Protocols for Apter AI Support. *Proceedings of the ACM on Human-Computer Interaction*, 9(7): 1–28.

Cabitza, F.; Campagner, A.; Ronzio, L.; Cameli, M.; Mandoli, G. E.; Pastore, M. C.; Sconfienza, L. M.; Folgado, D.; Barandas, M.; and Gamboa, H. 2023b. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, 138: 102506.

Cook, D. A.; and Triola, M. M. 2009. Virtual patients: a critical literature review and proposed next steps. *Medical education*, 43(4): 303–311.

Fogliato, R.; Chappidi, S.; Lungren, M.; Fisher, P.; Wilson, D.; Fitzke, M.; Parkinson, M.; Horvitz, E.; Inkpen, K.; and Nushi, B. 2022. Who goes first? Influences of human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1362–1374.

Harvey, N.; and Fischer, I. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational behavior and human decision processes*, 70(2): 117–133.

Iniesta, R. 2023. The human role to guarantee an ethical AI in healthcare: a five-facts approach. *AI and Ethics*, 1–13.

Klein, G. A. 2022. *Snapshots of the Mind*. MIT Press.

Li, B.; XU, F.-M.; Wang, W.; DENG, Z.-J.; and ZHANG, J.-W. 2010. Anchoring effects: types, influential factors and intervention measures. *Advances in Psychological Science*, 18(01): 34.

Natali, C.; Campagner, A.; and Cabitza, F. 2024. Answering the Call to Go Beyond Accuracy: An Online Tool for the Multidimensional Assessment of Decision Support Systems. In *BIOSTEC (2)*, 219–229.

Parasuraman, R.; and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors*, 39(2): 230–253.

Schmitt, A.; Wambsganss, T.; Söllner, M.; and Janson, A. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. In *ICIS*.

Soper, D. S. 2020. Informational social influence, belief perseverance, and conservatism bias in web interface design evaluations. *IEEE Access*, 8: 218765–218776.

Yaniv, I. 2004. Receiving other people's advice: Influence and benefit. *Organizational behavior and human decision processes*, 93(1): 1–13.