From Accuracy to Readiness: Metrics and Benchmarks for Human–AI Decision-Making

Min Hun Lee
mhlee@smu.edu.sg
Singapore Management University
Singapore

Abstract

Artificial intelligence (AI) systems are deployed as collaborators in human decision-making. Yet, evaluation practices focus primarily on model accuracy rather than whether human-AI teams are prepared to collaborate safely and effectively. Empirical evidence shows that many failures arise from miscalibrated reliance, including overuse when AI is wrong and underuse when it is helpful.

This paper proposes a measurement framework for evaluating human-AI decision-making centered on team readiness. We introduce a four-part taxonomy of evaluation metrics spanning outcomes, reliance behavior, safety signals, and learning over time, and connect these metrics to the Understand–Control–Improve (U–C–I) lifecycle of human-AI onboarding and collaboration.

By operationalizing evaluation through interaction traces rather than model properties or self-reported trust, our framework enables deployment-relevant assessment of calibration, error recovery, and governance. We aim to support more comparable benchmarks and cumulative research on human–AI readiness, advancing safer and more accountable human–AI collaboration.

CCS Concepts

• **Human-centered computing** → **HCI theory, concepts and models**; **HCI design and evaluation methods**; • **Computing methodologies** → *Artificial intelligence.*

Keywords

Human-Centered AI, Human-AI Collaboration, Human–AI Decision Making, Appropriate Reliance, AI Evaluation Metrics, AI Governance

ACM Reference Format:

Min Hun Lee. 2026. From Accuracy to Readiness: Metrics and Benchmarks for Human–AI Decision-Making. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3772363.3798377

1 Introduction

Artificial intelligence (AI) systems are increasingly deployed as collaborators rather than autonomous decision-makers, supporting human judgment in high-stakes domains such as healthcare [8, 28, 29, 46] and public services [23, 50]. In these settings, AI systems increasingly shape how people interpret evidence, calibrate confidence, allocate responsibility, and ultimately make decisions [10, 19, 24, 29].

Over the past several years, empirical Human–AI Interaction (HAI) research has demonstrated that model performance alone is insufficient for safe and effective human–AI collaboration: even highly accurate systems can yield worse human–AI outcomes when users follow incorrect advice, ignore correct advice, or apply inconsistent intervention strategies under uncertainty [4, 7, 11, 18, 24]. Complementing these findings, research on accountable and trustworthy AI emphasizes governance mechanisms, such as oversight, contestability, auditing, and responsibility across deployment [20, 30, 35, 36, 38, 43, 44]. Meanwhile, explainable AI (XAI) and interactive ML research has proposed many mechanisms—feature attributions, examples, counterfactuals, rules, and uncertainty estimates—to make model behavior intelligible [3, 12, 13, 16, 22, 39, 40, 47]. However, empirical evidence across HAI and XAI suggests these techniques do not reliably improve decision quality by default. Instead, their effects depend on task context, user expertise, timing, and interactions with human intuition and confidence [7, 11, 18, 24, 25].

Despite progress on mechanisms, evaluation practices remain misaligned with how human–AI systems fail in practice during real-world deployment. Many studies emphasize model accuracy, explanation fidelity, or self-reported trust [14, 17, 24, 40], implicitly assuming these proxies reflect whether users are ready to collaborate with AI safely and effectively. Yet, trust often poorly predicts reliance behavior, and explanations can increase overreliance by providing a false sense of certainty or legitimacy [4, 11, 14, 24, 25]. Consequently, real-world failures persist not only due to model error, but due to miscalibrated human reliance—overreliance when AI is wrong, underuse when AI is helpful, and brittle "local" adaptations that do not generalize across cases [7, 11, 14, 18, 25]. Critically, these failure modes are often invisible when evaluation reports only accuracy, perceived trust, or explanation satisfaction.

In this paper, we argue that resolving this gap requires shifting evaluation from "how good is the model?" to "how ready is the human–AI team?": whether users can recognize failures, calibrate reliance, and remain accountable under realistic constraints [4, 7, 11, 24]. We focus on onboarding, calibration, and governance as the early-deployment phase where reliance patterns are formed and where many downstream failures originate [9, 34]. Building on this direction, our work reframes onboarding as a measurable learning intervention organized around **Understand–Control–Improve (U–C–I)**, extending recent work on AI onboarding and explanation-supported learning for clinical decision-making [26, 27]. We treat

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
CHI EA '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2281-3/2026/04
https://doi.org/10.1145/3772363.3798377

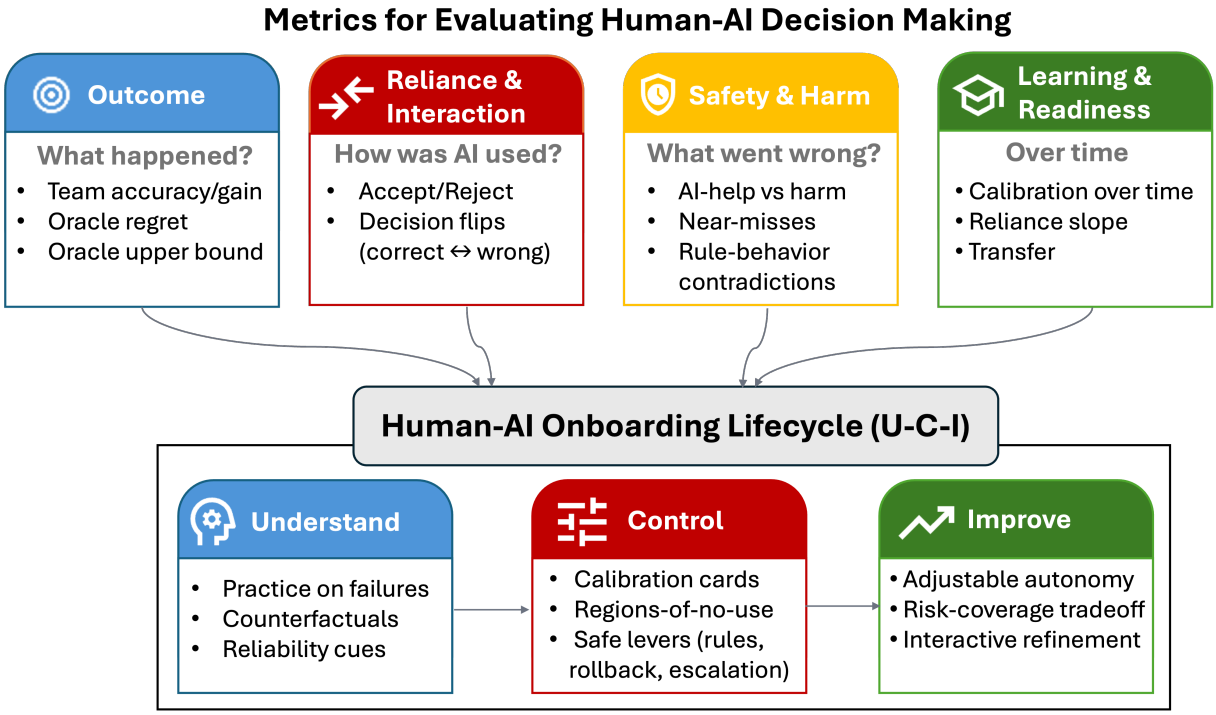

Figure 1: We propose a four-part taxonomy of metrics for human–AI onboarding (top) and show how each metric family becomes observable and actionable across the human-AI onboarding lifecycle (Understand–Control–Improve) (bottom).

onboarding broadly as the process through which users learn to work effectively with AI systems in real decision-making settings. In **Understand**, users develop mental models of model behavior, boundary conditions, and failure modes through structured practice on curated failure sets and counterfactual examples that reveal how small input changes can flip predictions [5, 25, 45]. In **Control**, users learn how to calibrate reliance and apply safe interventions using lightweight supports, such as *calibration cards*, artifacts that summarize when AI predictions are reliable or unreliable (e.g. "when to trust", "when to double-check") [30], common failure modes [9, 26], and recommended operating points (e.g. thresholds or escalation rules) [31], alongside regions-of-no-use, contexts where AI recommendations should not be trusted [27], and safe levers, user-facing controls (e.g. rule/threshold edits) that allow bounded intervention in AI behavior with preview, rollback, and audit trails to support contestability [1, 2, 21, 29, 30, 33] and accountability [12, 34, 38]. In **Improve**, teams iteratively refine collaboration strategies and governance policies using feedback from newly observed failures to update training content, thresholds, and governance practices [4, 21, 24, 29, 30].

We organize these measures around the Understand–Control–Improve (U–C–I) lifecycle. U–C–I describes when key capabilities in human–AI collaboration develop: users first learn model behavior and limitations (Understand), then calibrate how and when AI should be used in practice (Control), and finally refine collaboration strategies and governance policies over time (Improve). The four metric families describe what should be measured across this lifecycle: Outcome metrics evaluate decision quality, Reliance and Interaction metrics capture how AI advice is adopted or rejected, Safety and Harm metrics identify high-risk collaboration failures, and Learning and Readiness metrics measure how these behaviors evolve over repeated use. Together, the taxonomy makes the U–C–I lifecycle observable, enabling evaluation of how human–AI collaboration evolves over time.

Prior work has proposed a variety of measures for evaluating human–AI interaction, including trust, reliance, agreement with model predictions, and decision accuracy [6, 15, 24, 42]. However, these measures are often studied in isolation and therefore do not capture the full lifecycle of human–AI collaboration. We synthesize these existing constructs into four complementary metric families: outcome quality, reliance behavior, safety and harm signals, and learning over time. These categories reflect four practical questions that arise when deploying AI decision-support systems: What happened? How was the AI used? What went wrong? And how does collaboration evolve over time?

Building on this framing, we further specify how these metrics can be computed directly from observable interaction traces rather than inferred attitudes or model properties. Our taxonomy spans four complementary classes.

(1) Outcome metrics: capture decision quality beyond raw model accuracy, such as team gain and avoidable error (e.g. regret relative to the best achievable human–AI decision), reflecting whether AI involvement ultimately improves or degrades outcomes [4, 17].

(2) Reliance & Interaction metrics: characterize how AI advice shapes human judgments, including accept-on-wrong, changed-to-wrong, override frequency and timing, and reliance slope, which operationalize behavioral calibration and sensitivity to AI correctness [7, 11, 24, 25].
(3) Safety & Harm metrics: attribute risk to AI influence and governance breakdowns rather than human error alone, including AI-induced harm, near-misses, contradictions between rules and behavior, and rollback or escalation events [14, 38].
(4) Learning & Readiness metrics: assess whether onboarding produces durable skill, such as failure identification, explanation comprehension, and retention or transfer across cases, tasks, or model versions [9, 19].

These four metric families can be instantiated across a wide range of decision-support settings. For example, in a clinical triage system [31], outcome metrics measure the accuracy of the final human–AI decision, reliance metrics capture how often clinicians accept or override AI recommendations, safety metrics detect harmful deferrals to incorrect AI predictions, and learning metrics track how reliance evolves across repeated cases.

These metrics are not standalone statistics. They are computed from decision traces (e.g. accept, override, change), error attribution (AI-influenced versus independent errors), and learning signals (e.g. pre/post onboarding probes, time-to-calibration, cross-case transfer). As a result, each metric class maps naturally to stages of the **Understand–Control–Improve (U–C–I)** onboarding lifecycle, where it becomes both observable (during interaction) and actionable (through training, control levers, or governance interventions). This structure moves evaluation beyond accuracy and trust toward cumulative, deployment-relevant evidence of human–AI readiness.

This framework surfaces a measurement and benchmarking agenda for the human–AI interaction community:

- When does a user become "AI-ready"?
- Which reliance and harm metrics generalize across domains?
- How should governance be evaluated in use—beyond documentation—through behaviors such as contestation, rollback, escalation, and auditability?

Answering these questions would enable cumulative science and more deployment-relevant evidence for safe human–AI collaboration. We argue that progress in human–AI collaboration requires shifting from evaluating models in isolation to evaluating human–AI teams, and from reporting isolated metrics to developing benchmarkable measures of readiness, calibration, and governance.

Positioning: Our work complements prior frameworks for measuring reliance in human–AI systems [17] and surveys of human–AI decision-making metrics [24]. While these works catalog existing measures or analyze reliance behavior, we focus on evaluation during onboarding and early deployment, where reliance patterns are formed and many downstream failures originate. We therefore propose a structured taxonomy of evaluation metrics and map these metrics to actionable stages in the Understand–Control–Improve (U–C–I) lifecycle.

Contribution: We contribute a *unified, traced-based evaluation framework* for human–AI readiness:

- A metric taxonomy spanning outcomes, reliance, harm, learning
- Trace-based metric definitions grounded in interaction logs
- A mapping from metrics to actionable **U–C–I** design interventions (Tables 1–2; Appendix A).

2 Why Accuracy Alone Is Insufficient

2.1 Why Current Evaluation Fails

Despite rapid advances in model performance, many failures of human–AI systems arise after deployment, during everyday use in real workflows. A growing body of HAI research suggests that this gap is not primarily due to insufficient model accuracy, but to a mismatch between how systems are evaluated and how they are actually used [17, 24]. In practice, AI systems are embedded in time pressure, institutional norms, accountability structures, and evolving user strategies, which are rarely reflected in standard evaluation protocols. The following three evaluation assumptions illustrate this mismatch.

2.1.1 Accuracy ≠ Safety. Accuracy measures whether a model's prediction matches ground truth, but it does not capture the quality of human–AI decisions. In high-stakes settings, such as healthcare, multiple studies show that users may change initially correct judgments to incorrect ones after seeing AI advice, a phenomenon often referred to as AI-induced error or automation bias [4, 7, 25]. These errors are invisible in standard accuracy (e.g. AUROC, or F1 metrics), which treat AI outputs as independent of human behavior. Moreover, accuracy does not distinguish between errors that users recognize and recover from versus errors that propagate silently into downstream decisions, documentation, or treatment plans [7, 14]. As a result, systems that appear high-performing in offline benchmarks may still increase harm when integrated into real workflows where AI advice shapes human judgment.

2.1.2 Trust ≠ Reliance. Trust is frequently measured through post-task surveys or Likert-scale questionnaires, yet behavioral evidence consistently shows weak alignment between reported trust and actual reliance [7, 11, 24, 25]. Users may report low trust while still following AI recommendations under time pressure, cognitive load, or organizational expectations. Conversely, users may report high trust while selectively ignoring AI advice in critical or ambiguous cases [4, 25]. This disconnect arises because trust captures attitudes, whereas reliance reflects situated behavior under constraints—including workload, accountability, and perceived risk. Evaluations that rely primarily on trust scores therefore miss when, how, and why users defer to or override AI advice in practice, obscuring important safety and governance concerns.

2.1.3 Performance ≠ Readiness. High task performance during evaluation does not imply that users are prepared for real-world deployment. Short-term performance gains can mask brittle strategies, such as copying AI outputs without understanding underlying uncertainty or failure modes [7, 18, 25]. In contrast, readiness depends on whether users can recognize when AI is likely wrong, interpret confidence and uncertainty appropriately, and recover from errors when they occur [9, 19, 25, 37, 41]. These capacities

(e.g. failure detection, uncertainty interpretation, and error recovery) are rarely measured explicitly, yet they determine whether human–AI systems remain safe over time, under distribution shift, and as models or workflows evolve [14, 24].

Together, these gaps point to a fundamental mismatch: we often evaluate AI systems as artifacts optimized for predictive performance, but deploy them as teammates embedded in human workflows. Addressing this mismatch requires evaluation frameworks that capture not only what the AI predicts, but how humans learn to work with it [9], rely on it [17, 24], and govern it over time.

2.2 Reframing Onboarding as a Measurable Process

To address this mismatch, we reframe onboarding not as documentation, demos, or one-off training, but as a *measurable learning intervention* that prepares users to collaborate with AI safely in real workflows [9, 19]. Drawing on prior work in human–AI collaboration, explainable AI, learning-by-doing, and AI onboarding for clinical decision-making [7, 9, 11, 24, 26, 27], we conceptualize onboarding as the process through which users acquire durable skills for forming accurate mental models of AI reliability, calibrating reliance, and enacting accountability under realistic constraints.

Effective onboarding supports at least four competencies. First, users learn to **detect reliability boundaries**: when AI is likely correct or incorrect rather than assuming uniform performance across cases, contexts, or subpopulations [9]. Second, users learn to **calibrate reliance**, adjusting when to accept, question, or override AI advice based on evidence and uncertainty cues [4, 7, 11]. Third, users learn to **exercise safe control and contestability**, including how to intervene [16, 21], escalate ambiguous cases, and use rollback or audit mechanisms when AI advice conflicts with domain judgment or policy requirements [34, 35, 38]. Fourth, users learn to **understand delegation and autonomy**, recognizing how responsibility shifts between human and AI under different operating modes (e.g., decision support vs. selective deferral) and how these choices affect outcomes and accountability [4, 17, 19, 31, 49].

These abilities cannot be inferred from model properties or self-reported attitudes alone; they must be measured behaviorally through *interaction traces over time* (e.g. acceptance/override patterns, sensitivity to AI correctness, failure detection rates, and recovery actions across cases and changing conditions) [11, 19, 24].

2.3 A Taxonomy of Metrics for Human–AI Onboarding & Decision-Making

Building on empirical findings across healthcare AI onboarding, decision-support evaluation, uncertainty-aware delegation, and accountable AI systems, we propose a taxonomy of metrics that capture complementary aspects of onboarding and collaboration [9, 17, 24, 25, 37, 38]. Our taxonomy separates four evaluation questions: *what happened*, *how AI was used*, *what went wrong*, and *what changed over time*—dimensions often conflated or omitted in prior evaluations [7, 14, 24]. Full metric definitions and equations are provided in Appendix A.

2.3.1 Outcome Metrics (What happened?) Outcome metrics capture the quality of final human–AI decisions beyond raw model correctness, reflecting whether AI involvement ultimately improves or degrades task outcomes [4, 17]. We report: (i) **team gain** relative to human-only and AI-only baselines, and (ii) **regret_best**, which quantifies avoidable error relative to an oracle that selects the better of the initial human decision and AI prediction per case [17]. We further distinguish **error recovery vs. error amplification**, separating cases where AI helps users correct initial mistakes from cases where AI induces harm that would not otherwise occur [14, 18]. **Oracle best accuracy** is treated as a reference upper bound rather than an operational target, enabling diagnosis of collaboration failures distinct from model limitations [17] (Appendix A).

2.3.2 Reliance & Interaction Metrics (How was AI used?) Reliance metrics characterize *how* AI advice shapes human decisions, operationalizing behavioral calibration rather than subjective attitudes [7, 11, 24]. We track: (i) **accept-on-wrong** (agreeing with incorrect AI), (ii) **changed-to-wrong** (switching from a correct human judgment to an incorrect final decision after seeing AI), (iii) **override frequency and timing**, and (iv) **local vs. global update asymmetry** (i.e. whether users treat a failure as case-specific or revise their broader mental model of AI reliability) [32, 48]. These measures expose overreliance, underuse, and brittle strategies that are invisible in aggregate accuracy [4] (Appendix A).

2.3.3 Safety & Harm Metrics (What went wrong?) Safety metrics attribute harm to AI influence and governance breakdowns rather than human error alone [14, 35, 38]. We include: (i) **AI-harm** (cases where AI causes a correct initial human decision to become wrong), (ii) **near-misses** (high-risk disagreements narrowly avoided), and (iii) **governance-in-use signals** such as contradictions between rules and behavior, rollback events, and escalation actions. These metrics operationalize accountability as enacted behavior rather than documentation alone [34, 38] (Appendix A).

2.3.4 Learning & Onboarding Metrics (What changed over time?) Learning metrics assess whether onboarding produces durable skill [9] rather than transient performance gains. We measure: (i) **calibration gap** (confidence vs. correctness), (ii) **reliance slope** (acceptance sensitivity to AI correctness), (iii) **stability under distribution shift**, and (iv) **transfer** across tasks, cases, or model versions. These targets operationalize "AI readiness" as a behavioral capability that persists beyond a single evaluation outcome [7, 17, 24] (Appendix A).

In operational settings, many of these metrics can be computed directly from interaction logs that record initial human decisions, AI recommendations, and final outcomes. In large-scale deployments, collecting these signals may require event-logging infrastructure similar to observability pipelines used in production ML systems. When ground-truth labels are delayed or expensive, practitioners may estimate some metrics through sampling strategies or proxy signals such as disagreement events or escalation rates. In privacy-sensitive settings, behavioral traces should be collected with appropriate aggregation and anonymization mechanisms.

2.3.5 Calibration & Governance as First-Class Targets. Across domains, outcomes depend less on raw predictive accuracy and more on whether users *calibrate reliance*, accepting AI when it is likely correct and overriding it when it is likely wrong [4, 7, 24]. Even highly accurate systems can degrade team performance when users

over-rely on incorrect advice or fail to intervene at critical moments [7, 14, 24]. Thus, calibration should be treated as a *primary* evaluation target (e.g. accept-on-wrong, changed-to-wrong, reliance slope, calibration gap), not a byproduct of explainability.

Governance mechanisms (e.g. model cards, audit trails, policies) are necessary but insufficient on their own: accountability is enacted through everyday interaction, including how users contest AI, justify overrides, escalate cases, or rollback edits [30, 34, 35, 38]. Behavioral signals such as rollback frequency, escalation behavior, contradiction detection, and intervention latency provide empirical evidence of "governance in use," enabling assessment beyond documentation [14, 38].

2.3.6 Open Benchmarking Questions. Taken together, our framework raises foundational benchmarking questions for the HCI and HAI community:

- **When is a user "AI-ready"?** What behavioral criteria indicate readiness for deployment, beyond short-term task performance?
- **Which onboarding metrics generalize across domains?** Which measures of reliance, learning, and harm are robust to task context, expertise, and institutional setting?
- **How should governance mechanisms be evaluated empirically?** What behavioral signals best capture contestability, accountability, and safe intervention in use?
- **What should standardized human–AI benchmarks include beyond accuracy?** How can benchmarks reflect calibration, error recovery, and governance rather than prediction alone?

Addressing these questions is essential for cumulative, comparable, and deployment-relevant progress in human–AI collaboration research.

3 Discussion and Conclusion

This paper positions *measurement* rather than algorithmic novelty as a central bottleneck for safe and accountable AI deployment. By shifting evaluation toward calibration, learning, and governance, the proposed framework aims to support: (i) comparable evaluation across studies and domains, (ii) principled design of onboarding interventions grounded in learning outcomes, and (iii) policy-relevant assessment of AI governance as enacted in practice. In addition, we provide an agenda for future CHI workshops, surveys, benchmarks, and research programs focused on human–AI teaming rather than model-centric performance. **As a limitation, this taxonomy should be understood as a starting point rather than a finalized standard**: it synthesizes recurring measures and highlights gaps, and it will require community iteration, domain-specific validation, and refinement as new evidence and deployment contexts emerge. If we do not measure onboarding, calibration, and harm, we cannot claim that human–AI systems are ready for real-world collaboration. This work proposes a shared measurement agenda for evaluating human–AI teams—not as tools, but as socio-technical systems whose safety and effectiveness emerge through interaction over time. This framework provides a foundation for future evaluation protocols, benchmark design, and shared measurement standards for human–AI collaboration across domains.

Acknowledgments

This research was supported by the Resilient Workforces Institute, Singapore Management University, under the SMU Seed Fund (Grant ID: 2026-6026IR-25T040-SMUIRNYXXX), and by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (MOE-T2EP20223-0007). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

[1] Kars Alfrink, Ianus Keller, Gerd Kortuem, and Neelke Doorn. 2023. Contestable AI by design: Towards a framework. *Minds and Machines* 33, 4 (2023), 613–639.
[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
[3] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).
[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.
[5] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.
[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
[7] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 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 (2021), 1–21.
[8] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.
[9] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.
[10] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.
[11] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–32.
[12] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive model cards: A human-centered approach to model documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 427–439.
[13] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
[14] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health* 3, 11 (2021), e745–e750.
[15] Ben Green and Yiling Chen. 2019. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*. 90–99.
[16] Lijie Guo, Elizabeth M Daly, Oznur Alkan, Massimiliano Mattetti, Owen Cornec, and Bart Knijnenburg. 2022. Building trust in interactive machine learning via user contributed interpretable rules. In *Proceedings of the 27th international conference on intelligent user interfaces*. 537–548.
[17] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 221–236.

[18] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[19] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–20.

[20] Davinder Kaur, Suleyman Uslu, Kaley J Rittichier, and Arjan Durresi. 2022. Trustworthy artificial intelligence: a review. *ACM computing surveys (CSUR)* 55, 2 (2022), 1–38.

[21] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[22] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Amy J Ko, and Ian Oberst. 2011. Why-oriented end-user debugging of naive Bayes text classification. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 1 (2011), 1–31.

[23] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[24] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[25] Min Hun Lee and Chong Jun Chew. 2023. Understanding the Effect of Counterfactual Explanations on Trust and Reliance on AI for Human-AI Collaborative Clinical Decision Making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.

[26] Min Hun Lee, Silvana Xin Yi Choo, Shamala D Thilarajah, et al. 2024. Improving health professionals' onboarding with ai and xai for trustworthy human-ai collaborative decision making. *arXiv preprint arXiv:2405.16424* (2024).

[27] Min Hun Lee, Renee Bao Xuan Ng, Silvana Xinyi Choo, and Shamala Thilarajah. 2024. Interactive Example-based Explanations to Improve Health Professionals' Onboarding with AI for Human-AI Collaborative Decision Making. *arXiv preprint arXiv:2409.15814* (2024).

[28] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-design and evaluation of an intelligent decision support system for stroke rehabilitation assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[29] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[30] Min Hun Lee and Justin Yu Feng Teo. 2026. RuleEdit: Failure-Guided Human–AI Model Editing with Prospective Impact Preview. (2026).

[31] Min Hun Lee and Martyn Zhe Yu Tok. 2025. Towards uncertainty aware task delegation and human-ai collaborative decision-making. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2274–2289.

[32] Q Vera Liao and Jennifer Wortman Vaughan. 2023. Ai transparency in the age of llms: A human-centered research roadmap. *arXiv preprint arXiv:2306.01941* 10 (2023).

[33] Henrietta Lyons, Eduardo Velloso, and Tim Miller. 2021. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[35] Jakob Mökander and Luciano Floridi. 2022. From algorithmic accountability to digital governance. *Nature Machine Intelligence* 4, 6 (2022), 508–509.

[36] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. 2024. Accountability in artificial intelligence: What it is and how it works. *Ai & Society* 39, 4 (2024), 1871–1882.

[37] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q Vera Liao, and Nikola Banovic. 2023. Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th international conference on intelligent user interfaces*. 379–396.

[38] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[40] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[41] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 410–422.

[43] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gonzalez Zelaya, and Aad Van Moorsel. 2020. The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 272–283.

[44] Kush R Vashney. 2022. *Trustworthy machine learning*. Independently published.

[45] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E Hines, John P Dickerson, and Chirag Shah. 2020. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).

[46] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–18.

[47] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[48] Xinru Wang and Ming Yin. 2023. Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[49] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582* (2020).

[50] Aleš Završnik. 2020. Criminal justice, artificial intelligence systems, and human rights. In *ERA forum*, Vol. 20. Springer, 567–583.

A Appendix: Detailed Metrics (Organized by the Four-Part Taxonomy)

Let $\{(y_j, h_{0j}, a_j, h_{1j}, c_j, t_j)\}_{j=1}^{N}$ denote N decision instances, where y_j is the ground truth, h_{0j} the participant's initial decision, a_j the AI prediction, h_{1j} the participant's final decision after viewing AI output, c_j the participant's reported confidence (if available), and t_j timestamps or interaction events recorded in system logs.

We organize metrics by *what happened*, *how AI was used*, *what went wrong*, and *what changed over time*, following prior analyses of human–AI reliance and collaboration behavior [7, 17, 24, 25].

A.1 Outcome Metrics (What happened?)

Accuracies and team gains. These metrics describe decision quality at the human, AI, and human–team level.

- **Human accuracy** (Acc_{h0}): proportion of cases correctly solved by the human before seeing AI.

$$\mathrm{Acc}_{h0} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{0j} = y_j]$$

- **AI accuracy** (Acc_{ai}): proportion of cases correctly predicted by the AI.

$$\mathrm{Acc}_{ai} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[a_j = y_j]$$

- **Team accuracy** (Acc_{team}): proportion of cases where the final human–AI decision is correct.

$$\mathrm{Acc}_{team} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{1j} = y_j]$$

- **TeamGain vs. Human**: improvement (or degradation) from human-only decisions to AI-assisted decisions.

$$\mathrm{Acc}_{team} - \mathrm{Acc}_{h0}$$

- **TeamGain vs. AI**: improvement (or degradation) from AI-only predictions to the final team decision.

$$\mathrm{Acc}_{team} - \mathrm{Acc}_{ai}$$

Oracle upper bound and regret. These metrics separate model limitations from collaboration failures. Define the oracle-correct indicator as:

$$\mathrm{Oracle}_j = \mathbb{I}\big[(h_{0j} = y_j) \vee (a_j = y_j)\big].$$

- **Oracle best accuracy** (Acc_{oracle}): upper bound on achievable team performance if one always selected the correct agent.

$$\mathrm{Acc}_{oracle} = \frac{1}{N}\sum_{j=1}^{N}\mathrm{Oracle}_j$$

- **Regret_best**: proportion of avoidable errors where the team fails despite at least one agent being correct [17].

$$\frac{1}{N}\sum_{j=1}^{N}\left(\mathrm{Oracle}_j - \mathbb{I}[h_{1j} = y_j]\right)$$

Error recovery vs. error amplification (derived outcome effects). Beyond aggregate accuracy and regret, we distinguish whether AI involvement helps users recover from errors or amplifies them. *Error recovery* refers to cases where an initially incorrect human decision becomes correct after viewing AI output, while *error amplification* refers to cases where a correct initial human decision becomes incorrect due to AI influence. These outcome-level effects are not captured by accuracy or regret alone, but are critical for assessing whether AI improves or degrades real-world decision quality. We operationalize error recovery and amplification through the help–harm decomposition (AI-help vs. AI-harm) and complementary decision-change metrics (ChangedToRight vs. ChangedToWrong), defined in subsequent sections of the appendix.

A.2 Reliance & Interaction Metrics (How was AI used?)

Reliance conditioned on AI correctness. These metrics capture behavioral reliance patterns, including appropriate reliance as well as over- and under-reliance.

Let $C = \{j : a_j = y_j\}$ and $\mathcal{W} = \{j : a_j \neq y_j\}$.

- **Accept-on-correct**: tendency to follow AI when it is correct.

$$\Pr(h_1 = a \mid a = y) = \frac{1}{|C|}\sum_{j\in C}\mathbb{I}[h_{1j} = a_j]$$

- **Reject-on-wrong**: ability to reject incorrect AI advice.

$$\Pr(h_1 \neq a \mid a \neq y) = \frac{1}{|\mathcal{W}|}\sum_{j\in\mathcal{W}}\mathbb{I}[h_{1j} \neq a_j]$$

- **Reject-on-correct**: unnecessary rejection of correct AI advice.

$$\Pr(h_1 \neq a \mid a = y)$$

- **Accept-on-wrong**: overreliance on incorrect AI predictions.

$$\Pr(h_1 = a \mid a \neq y)$$

Decision-change behaviors. These metrics distinguish beneficial from harmful decision updates.

- **Changed**: proportion of cases where the participant changes an initial decision after seeing AI.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{1j} \neq h_{0j}]$$

- **ChangedToRight**: beneficial changes from incorrect to correct.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{1j} \neq h_{0j}]\mathbb{I}[h_{0j} \neq y_j]\mathbb{I}[h_{1j} = y_j]$$

- **ChangedToWrong**: harmful changes induced after seeing AI.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{1j} \neq h_{0j}]\mathbb{I}[h_{0j} = y_j]\mathbb{I}[h_{1j} \neq y_j]$$

Calibration and timing.

- **Reliance slope**: sensitivity of reliance to AI correctness; higher values indicate better calibration [7].

$$\Pr(h_1 = a \mid a = y) - \Pr(h_1 = a \mid a \neq y)$$

- **Intervention latency**: average time taken to confirm or override AI, reflecting hesitation and contesting AI recommendations.

$$\frac{1}{N}\sum_{j=1}^{N}(t_j^{\text{confirm/override}} - t_j^{\text{AI}})$$

Local vs. global update asymmetry. To distinguish case-specific reactions from durable belief updates about AI reliability, we measure whether responses to an AI failure generalize beyond the current instance. *Local updates* are reflected by isolated overrides or rejections confined to the current case, whereas *global updates* manifest as systematic changes in reliance behavior on subsequent cases. We operationalize global updating by comparing reliance metrics (e.g., accept-on-wrong, reliance slope) before versus after observed AI failures, and quantify update asymmetry as the degree to which behavior changes persist across subsequent cases rather than reverting immediately. This framing aligns with evidence that user behavior can shift under model/explanation updates and with calls to study transparency via user mental models over time [32, 48].

A.3 Safety & Harm Metrics (What went wrong?)

Help–harm decomposition. These metrics attribute outcome changes to AI influence rather than overall accuracy alone.

- **AI-help**: cases where AI corrects an initially wrong human decision.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{0j} \neq y_j \wedge h_{1j} = y_j]$$

- **AI-harm**: cases where AI causes a correct human decision to become wrong.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{0j} = y_j \wedge h_{1j} \neq y_j]$$

- **Missed-help**: failures to adopt correct AI advice when the human is wrong.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{0j} \neq y_j \wedge a_j = y_j \wedge h_{1j} \neq a_j]$$

- **Correct-ignore**: appropriate rejection of incorrect AI advice.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[h_{0j} = y_j \wedge a_j \neq y_j \wedge h_{1j} \neq a_j]$$

- **Near-miss rate**: proportion of high-risk cases where an incorrect AI recommendation was narrowly avoided through human intervention or override.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[a_j \neq y_j \wedge h_{1j} = y_j \wedge \text{Risk}_j = \text{high}]$$

Governance-in-use signals. These metrics operationalize governance as observable behavior in practice rather than documentation. Let R_j indicate a rollback, E_j an escalation, and π_j a policy rule for case j.

- **Rollback rate**: frequency of reversing AI-influenced decisions after review.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[R_j = 1]$$

- **Escalation rate**: proportion of cases referred for human or institutional oversight.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[E_j = 1]$$

- **Rule–behavior contradiction**: violations where required actions (e.g., escalation) are not taken.

$$\frac{1}{N}\sum_{j=1}^{N}\mathbb{I}[\pi_j = \text{escalate} \wedge E_j = 0]$$

A.4 Learning & Readiness Metrics (What changed over time?)

These metrics assess whether onboarding produces durable, transferable user capability.

- **Calibration gap**: misalignment between reported confidence and actual correctness.

$$\frac{1}{N}\sum_{j=1}^{N}|c_j - \mathbb{I}[h_{1j} = y_j]|$$

- **Reliance slope over time (behavioral calibration)**: change in sensitivity of agreement to AI correctness across sessions/blocks. For block/session k:

$$\text{Slope}_k = \Pr(h_1 = a \mid a = y)_k - \Pr(h_1 = a \mid a \neq y)_k$$

and learning can be summarized as $\Delta\text{Slope} = \text{Slope}_{post} - \text{Slope}_{pre}$

- **Retention**: stability of calibration-related metrics across multiple sessions or time intervals.

$$|\text{Calib}_{\text{session } k} - \text{Calib}_{\text{session } k+1}|$$

- **Transfer**: consistency of performance or reliance across tasks, datasets, or model versions.

$$|\text{Metric}_{\text{task A}} - \text{Metric}_{\text{task B}}|$$

Table 1: Mapping metrics to observable data sources, U–C–I onboarding stages, and the corresponding design actions they enable. (Part 1)

Metric (example)	Data source (trace)	U–C–I stage	Design action (what you do when it's bad)
Outcome metrics (What happened?)			
Team accuracy / TeamGain	Final decision h_1, ground truth y; condition logs	Improve	Adjust delegation policy (when to defer / override); revise workflow to route high-risk cases to human review.
Regret_best / Oracle gap	h_0, a, h_1, y per case	Improve	Diagnose collaboration failures (not model limits); target training on cases where either human or AI was correct but the team failed.
Error recovery vs. amplification	$h_0 \rightarrow h_1$ transitions + y	Improve	Identify whether UI causes harmful flips; refine prompts/explanations to reduce changed-to-wrong; add guardrails for high-stakes edits.
Reliance & interaction metrics (How was AI used?)			
Accept-on-wrong **Reject-on-wrong**	Agreement with AI ($h_1 = a$) conditioned on AI correctness	Understand + Control	Curate "failure sets"; add reliability cues; introduce regions-of-no-use; require justification or second-check for high-risk acceptance.
ChangedToWrong **ChangedToRight**	Decision-change events ($h_1 \neq h_0$) + y	Understand + Control	Refine onboarding tasks to expose boundary conditions; add counterfactual practice; redesign explanation timing to prevent harmful flips.
Override rate + timing	Override events + timestamps t_j	Control	Add safe levers (sandbox preview, rollback); reduce friction for appropriate overrides; introduce escalation shortcuts for uncertain cases.
Reliance slope	$\Pr(h_1 = a \mid a = y) - \Pr(h_1 = a \mid a \neq y)$	Understand	Diagnose calibration; if low, strengthen training on discriminating correct vs. incorrect AI; improve uncertainty communication.
Intervention latency	$(t^{confirm/override} - t^{AI})$	Control	Tune interaction costs; add "pause-and-check" for risky cases; streamline override/escalation to reduce delayed intervention.

Table 2: Mapping metrics to observable data sources, U–C–I onboarding stages, and the corresponding design actions they enable. (Part 2)

Metric (example)	Data source (trace)	U–C–I stage	Design action (recommended response)
Safety & harm metrics (What went wrong?)			
AI-harm / AI-help	Help–harm decomposition from h_0, h_1, a, y	Control + Improve	Add guardrails where AI induces harm; adjust autonomy (e.g., limit deferral) in high-harm regions; prioritize model fixes for harm-heavy slices.
Missed-help / Under-reliance	$h_0 \neq y$, $a = y$, but $h_1 \neq a$	Understand	Improve "when to trust" instruction; show exemplars of correct AI behavior; add calibrated confidence cues for beneficial reliance.
Near-misses (high-risk disagreements)	High-stakes disagreement logs; risk labels; margin/uncertainty if available	Control	Trigger required second review; add risk-based escalation rules; refine "regions-of-no-use" policy.
Rule–behavior contradiction rate	Policy label π_j + behavior event (E_j)	Control + Improve	Fix workflow compliance gaps; redesign the UI to make required actions salient; update governance policy or training based on observed violations.
Rollback rate / Escalation rate	Rollback logs R_j; escalation logs E_j	Control + Improve	Audit contested decisions; improve contestability pathways; adjust accountability and when rollback is encouraged.
Learning & readiness metrics (What changed over time?)			
Calibration gap	Confidence c_j + correctness $\mathbb{I}[h_1 = y]$	Understand	Improve calibration cards; add feedback on confidence miscalibration; emphasize boundary conditions and failure modes.
Retention	Same metrics across sessions (pre/post; follow-up)	Improve	Iterate the onboarding curriculum; schedule refreshers; adapt materials to failure modes that "do not stick."
Transfer (across tasks/versions)	Metrics across task A vs. B, or model version v vs. v'	Improve	Update onboarding for new model versions; add regression tests for reliance and harm; retrain users on newly emerging failures.
Time-to-calibration	Rolling-window estimates of reliance slope / accept-on-wrong over time	Understand + Improve	Personalize onboarding length; stop training when stable calibration is achieved; allocate extra practice for slow-to-calibrate users.