

Wrong Together Is an Item-Level Quantity: Why Per-Person Reliance Measures Cannot Say Whether AI Assistance Correlates Human Errors, Why the Direct Probes Disagree, and the Reanalysis That Would Settle It

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Research on reliance on AI advice scores one decision-maker at a time: how often a person accepts correct advice, how often they reject wrong advice, and what that does to their accuracy. Deployment places many decision-makers in front of the same few models, and the theory of algorithmic monoculture warns that their remaining errors may then fall on the same cases, which erodes second opinions, committees and crowds even while each person improves. In this analysis the author reviews and synthesises published results to show that the two literatures cannot currently be joined, and why. Under two stated assumptions, exchangeable decision-makers and errors that are independent given the case, the pairwise covariance of two people's errors equals the variance, across cases, of the probability of erring. That variance splits into a part that per-person reliance rates do identify, set by the AI's error rate and by the gap between a person's error rate when the AI is wrong and when it is right, and parts that they cannot identify, because the standard metrics pool over cases and discard which cases a person got wrong. The direct probes disagree in the way this predicts. In a lung-cancer study where novices followed wrong AI advice about a third of the time, shared errors fell under assistance; in a mammography study where wrong suggestions were planted and most readers erred on those cases, the identifiable part alone puts the error correlation within reader groups at no less than roughly 0.14 to 0.32. Neither side has reported a scale-free, accuracy-adjusted correlation for aided against unaided decisions on the same cases; the nearest quantity, a committee dependence penalty, is not on a correlation scale. The paper gives a decision procedure for when correlation costs anything, and a reanalysis agenda for datasets that already hold the per-case records needed.

Keywords: human-AI decision making; appropriate reliance; algorithmic monoculture; error correlation; wisdom of crowds; outcome homogenization; automation bias; collective intelligence

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The empirical study of human reliance on AI advice is now large and methodologically mature. It began in human factors, where automation misuse and disuse were catalogued long before machine learning (Parasuraman & Riley, 1997) and where the design goal was named appropriate reliance rather than maximal trust (Lee & See, 2004). It continued through clinical decision support, where a systematic review of automation bias collected 74 studies of over-reliance and its mediators (Goddard et al., 2012), and through the judgement literature, which documented aversion to algorithms after people see them err, even when the algorithm outperforms the human (Dietvorst et al., 2015). In machine learning and human-computer interaction it became a field of its own, with a survey of the design choices in more than 100 papers (Lai et al., 2021), named phenomena such as explanations raising acceptance regardless of correctness (Bansal et al., 2021), interventions that reduce over-reliance (Bucinca et al., 2021; Vasconcelos et al., 2023), measurement proposals that split reliance into separate dimensions (Schemmer et al., 2023) or ground it in decision theory (Guo et al., 2024), and a meta-analysis of more than 100 experiments (Vaccaro et al., 2024).

In almost all of this work the unit of analysis is one person deciding one case, summarised over that person's cases. The outcome is that person's accuracy, or a ratio of their correct switches to their opportunities to switch. Deployment does not look like that. A hospital, a court system, a hiring market or a content-moderation team puts many decision-makers in front of the same model, or the same few models, on overlapping streams of cases. A second literature, on algorithmic monoculture, starts from exactly that fact. Kleinberg and Raghavan (2021) show that when firms converge on one ranking algorithm, total welfare can fall even when the shared algorithm is more accurate than each firm's own evaluator. Bommasani et al. (2022) introduce outcome homogenization, the extent to which the same individuals receive negative outcomes from every decision-maker, and find that sharing training data reliably increases it. Toups et al. (2023) report a trend toward systemic failure across deployed machine-learning systems. Kim et al. (2025), across more than 350 language models, find that errors are strongly correlated and that more accurate models are more correlated with each other.

The two literatures meet at an obvious question, which is the title of this paper's topic: when an intervention makes each person right more often, does it make people wrong together? A mechanical argument says yes. If

each assisted judgement is partly the person's own signal and partly the model's output, then raising reliance raises the share of every judgement that comes from one shared source, and the shared source's mistakes become everyone's mistakes. Fugener et al. (2021) report a multi-method study in which humans advised by an AI converged toward similar answers, individually better, while their unique knowledge shrank, and in which simulated crowds of assisted humans did far worse than crowds of unassisted ones. A recent formal model makes the same point as an individual-team reliance paradox, in which the reliance level that minimises each person's error can raise the error of the aggregated team decision (Tan, 2026).

The empirical probes that test the mechanism on humans do not line up behind it. In two pre-registered field experiments with medical novices diagnosing lung cancer from CT scans, the conditions in which participants used AI input in practice showed fewer shared errors than the unassisted control, and synthetic three-person committees in those conditions came closer to the independent-error benchmark (He et al., 2025). In a forecasting experiment with 991 participants, access to a language-model assistant raised individual accuracy, and the authors found no consistent effect on aggregate crowd accuracy: their pre-registered and exploratory analyses pointed in opposite directions (Schoenegger et al., 2024a). Meanwhile, the probes that do find convergence measure something else: the content of stories and ideas becomes more similar under generative assistance (Doshi & Hauser, 2024; Anderson et al., 2024; Padmakumar & He, 2024), which is homogenisation of output, not correlation of errors. And a deflationary philosophical analysis concludes that "monoculture is less problematic than its critics have supposed", partly because the independence that polyculture is supposed to supply was never there (Hedden & Raghavan, 2026).

In this analysis the author reviews and synthesises published results to show that this disagreement is not, at present, a disagreement about the world. It is the product of a missing measurement. The argument has four steps.

First, the paper derives what the bridge between the two literatures would have to be. Under two stated assumptions (decision-makers who are exchangeable, and errors that are independent given the case), the covariance between two people's errors is the variance across cases of the probability of erring. That is a standard identity, but it has a consequence for reliance research that has not been drawn out: the population-level quantity is a property of how error probability is distributed over cases, and the per-person reliance metrics in current use pool over cases before they are reported. They therefore cannot recover it in general.

Second, the variance splits by the AI's correctness into a between-class part and two within-class parts. The between-class part is identifiable from three numbers that reliance studies sometimes report: the AI's error rate, a person's error rate when the AI is wrong, and

their error rate when the AI is right. The within-class parts, which carry how strongly people agree on which wrong suggestions to follow and which correct suggestions to reject, are not identifiable from any pooled rate. The decomposition gives a lower bound on assisted error correlation from published reliance rates, and says exactly what extra record would turn the bound into an estimate.

Third, applied to the probes, the decomposition explains why they disagree. In He et al. (2025) novices diverged from wrong AI advice about two times in three, so the identifiable part is small. In Dratsch et al. (2023), where an AI suggestion was deliberately wrong on 12 of 40 mammograms and readers' accuracy on those cases fell to between 19.8% and 45.5%, the identifiable part alone implies, under the model's assumptions, an error correlation of at least roughly 0.14 to 0.32 depending on reader experience. Whether assistance raises correlation relative to the unassisted baseline then depends on a further quantity, the overlap between the cases that humans find hard and the cases the AI gets wrong, which the probes do not report either.

Fourth, the paper asks when correlation costs anything at all. It costs nothing to accuracy when each case is decided once by one person; it matters for second opinions, committees and crowds, and for an applicant who faces many decision-makers. The paper gives a decision procedure (Algorithm 1) that makes those distinctions explicit, and a reanalysis agenda. Several published datasets already contain the per-case, per-person records the missing measurement needs, with aided and unaided arms; a ten-dataset comparison of AI advice with an independence-preserving protocol was built from them (Berger et al., 2026). The measurement that would settle the question can be run on existing data.

The paper is a research agenda. Section 2 sets out what each literature measures. Section 3 states the model and the decomposition. Section 4 consolidates the direct probes in Table 1 and Figure 1. Section 5 explains their disagreement through four confounds. Section 6 asks which consumers of independence are harmed by correlation. Section 7 gives Algorithm 1. Section 8 states the reanalyses that would settle the open parts. Sections 9 to 12 take objections, place the paper against its nearest neighbours, list what is not known, and conclude.

2. Two literatures and what each one measures

2.1 Reliance research measures a person, pooled over cases

The reliance literature inherited its central concept from human factors. Parasuraman and Riley (1997) separated use, misuse (over-reliance), disuse (under-use, often after false alarms) and abuse of automation, and observed that large individual differences make use hard to predict. Lee and See (2004) recast the goal as trust that is calibrated to the automation's actual capability, so that

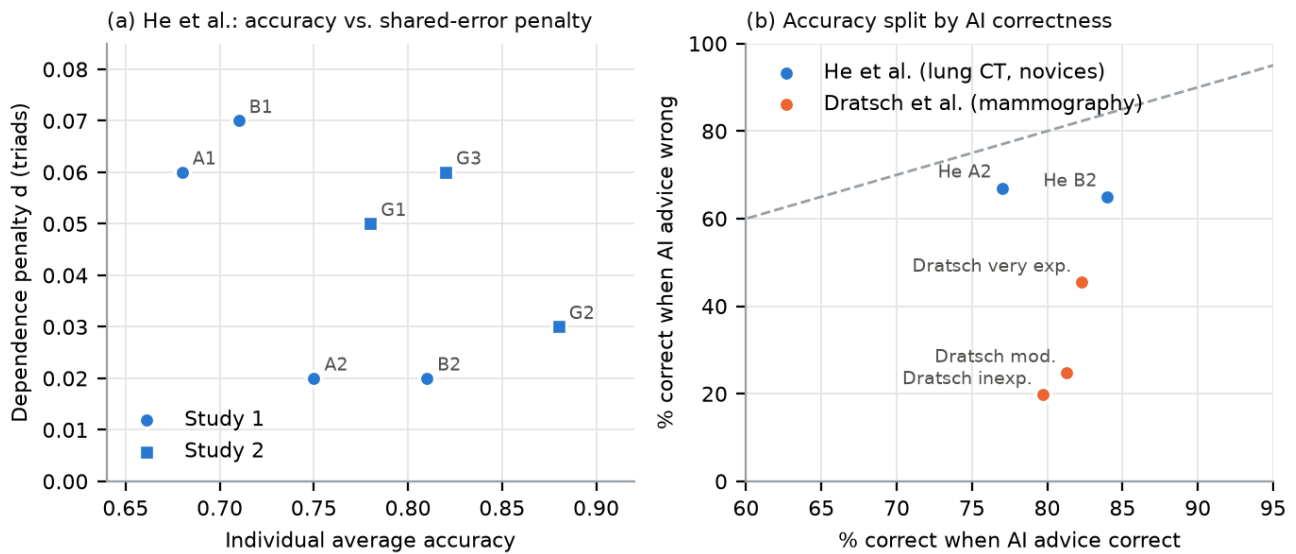


Figure 1. (a) Individual average accuracy against the dependence penalty d of synthetic three-person committees, for the seven conditions of He et al. (2025, Table 2); A1 and G1 are unassisted controls. (b) Percentage correct when the AI suggestion was correct against when it was wrong: He et al. (2025) Study 1 groups A2 and B2 (from the reported rates of diverging from correct and incorrect AI advice), and Dratsch et al. (2023) by reader experience. The dashed line is equal accuracy; the vertical distance below it is the gap that drives the between-class term. All values are as published by the cited studies.

reliance is appropriate case by case. Both framings are about a single operator and a single automated aid. Neither has a term for what happens when many operators share one aid.

The machine-learning era added measurement. Zhang et al. (2020) studied how displayed confidence and explanations affect accuracy and trust calibration when humans and an AI perform comparably. Bansal et al. (2021) asked whether explanations produce complementary performance, team accuracy above both human and AI alone, and observed that explanations raised acceptance of the AI's recommendation whether or not it was right. Bucinca et al. (2021), in an experiment with 199 participants, showed that cognitive forcing designs reduced over-reliance relative to simple explainable-AI designs. Vasconcelos et al. (2023) formalised the choice to engage with an explanation as a cost-benefit decision and showed that explanations can reduce over-reliance when engagement is cheap enough relative to its payoff. Lai et al. (2021) surveyed the decision tasks, AI assistance elements and evaluation metrics in more than 100 such papers. Steyvers and Kumar (2023) set out three challenges for the field: understanding the conditions for complementarity, assessing human mental models of the AI, and understanding the design choices of the interaction.

Two measurement proposals are worth stating precisely, because their structure is the obstacle this paper identifies. Schemmer et al. (2023) define Appropriateness of Reliance as a pair. Relative AI reliance (RAIR) is the number of cases in which a person's initial answer was wrong, the AI's advice was right, and the person switched to it, divided by the number of cases in which the initial answer was wrong and the advice right. Relative self-reliance (RSR) is the

analogous ratio for keeping a correct initial answer against wrong advice. Both are sums over cases in the numerator and denominator; the identity of the cases does not survive. Guo et al. (2024) argue that this definition lacks statistical grounding and propose instead a decision-theoretic definition, in which reliance is the probability of following the AI's recommendation and is separated from the person's difficulty in telling good recommendations from bad ones, with a rational benchmark for comparison. Their reliance is again a property of the person, averaged over the distribution of cases. That is the right object for the questions those papers ask. It is the wrong object for the question asked here.

A pair of medical studies shows that even the person-level picture is heterogeneous. Yu et al. (2024), studying 140 radiologists across 15 chest X-ray tasks, report that experience, subspecialty and familiarity with AI tools failed to predict who benefited from assistance, that lower performers did not consistently gain more, and that inaccurate AI predictions harmed performance on the aggregate of pathologies. Agarwal et al. (2023) find that radiologists do not capture the potential gains from AI assistance because they deviate from Bayesian updating: they partly underweight the AI's information and do not account for the correlation between their own information and the AI's prediction. That second finding is a person-level failure to model correlation between one human and one AI. The population-level analogue, correlation between humans induced by a shared AI, is not measured in either study.

The meta-analytic summary is sobering on its own terms. Across more than 100 experiments and over 300 effect sizes, human-AI combinations performed worse on average than the better of the human or the AI alone,

with losses concentrated in decision tasks and gains in content creation, and with gains when the human outperformed the AI and losses when the AI outperformed the human (Vaccaro et al., 2024). None of the outcomes pooled there is a between-person quantity.

2.2 Monoculture research measures a population, but of algorithms

The monoculture literature asks the population question, but its decision-makers are usually algorithms, and adoption is all or nothing. In the model of Kleinberg and Raghavan (2021), two firms hire in random order from a common pool. Each can use a private human evaluator, independent of the other firm's, or a shared algorithmic ranking that is identical for both. Under two conditions on the candidate distribution and the noise model, which hold for some parameterisations of random utility models and always for the Mallows model, they show that for any human accuracy there is a strictly higher algorithmic accuracy at which both firms rationally adopt the algorithm while total welfare would be higher had both kept their human evaluators. The mechanism is the loss of independent draws, not a shock. Peng and Garg (2024) extend the question to a two-sided matching market with many participants and find that, all else equal, monoculture selects less-preferred applicants when noise is well-behaved, while matching more applicants to their first choice.

Bommasani et al. (2022) move from welfare to the people decided about. Their component-sharing hypothesis says that decision-makers who share training data or models produce more homogeneous outcomes; on fairness benchmarks they find that shared training data reliably increases homogenization, while for shared foundation models the result depends on the adaptation method. Toups et al. (2023) apply ecosystem-level analysis to deployed machine-learning systems in three modalities and eleven datasets and report a consistent tendency toward systemic failure, in which some users are misclassified by every available system. Creel and Hellman (2022) argue that arbitrariness in a single algorithmic decision is not generally a moral wrong, but that the same algorithm, or algorithms built on the same data, applied across many contexts can exclude a person from a whole range of opportunities, which is. Jain et al. (2024) build on that point with a structural argument for algorithmic pluralism, in which no set of algorithms constitutes a severe bottleneck. Marx et al. (2020) show the other face of the same coin: prediction problems often admit competing models of near-equal accuracy whose individual predictions conflict, so a choice among them is partly arbitrary for the person decided about.

The language-model era has made the correlation measurable at scale. Kim et al. (2025) evaluate over 350 models on two leaderboards and a resume-screening task. On the Helm leaderboard, pairs of models choose the same wrong answer about 60% of the time when both are wrong, against a chance rate of one in three; even after controlling for shared provider, architecture and size, more accurate models have more correlated

errors. Goel et al. (2025) propose an accuracy-adjusted similarity metric, chance-adjusted probabilistic agreement, built on the observation that two accurate models agree often simply because both are usually right; they report that model mistakes are becoming more similar as capability rises. At the level of output rather than error, Wu et al. (2024) document generative monoculture, a narrowing of output diversity relative to the training data, and Bommasani et al. (2021) had named homogenization as a defining property of foundation models, since the defects of one base model are inherited by everything adapted from it.

What this literature does not contain is a human in the loop who partly relies. In Kleinberg and Raghavan's model a firm either uses the algorithm or does not. Peng and Garg's evaluators are algorithms with independent noise. Kim et al. measure model against model. The reliance literature's central empirical fact, that people follow some advice and not other advice, and that the pattern depends on the case, has no place in these models.

2.3 The collective-judgement literature already has the bridge, for human advisers

A third literature, older than both, has the missing piece in a different setting. Condorcet's jury theorem says that a majority of independent voters, each better than chance, is more likely to be right than any one of them. Ladha (1992) generalised the theorem to correlated votes and found that for large groups the conclusion survives under fairly general conditions, while for small groups the conditions on correlation are severe. Hong and Page (2004) showed in a model of problem solvers that a randomly drawn team can outperform a team of the individually best, because the best become similar to one another. Davis-Stober et al. (2014) formalised when a crowd is wise and found crowd wisdom robust even to biased and correlated judges, while arguing that what maximises it is not independence as such but judgements that differ systematically. Broomell and Budescu (2009) derived the correlation between two judges from the structure of the environment and the judges' cue use, and found the correlation among the cues to be the central driver: two judges who weight the same cues in the same environment will correlate however independently they work.

Two empirical results from this literature are directly relevant. Lorenz et al. (2011), in an experiment with 144 participants, showed that even mild social influence, seeing the average or the full set of others' estimates, narrowed the diversity of estimates without improving the collective error, and raised confidence. Kurvers et al. (2016), simulating groups from more than 20,000 real diagnoses by more than 140 doctors in breast and skin cancer, found that pooling independent judgements beats the best doctor in a group when the doctors' accuracies are similar but not when they differ too much. Soll and Larrick (2009) studied how people revise a judgement after seeing one other person's, and found that people favour choosing one of the two estimates where

averaging would usually have been more accurate. Vodrahalli et al. (2022), with more than 1,100 crowdworkers, found that whether people heeded advice depended on their beliefs about human versus AI performance on the task, and that once heeded, AI advice and equivalent advice from a group of peers were used similarly. An AI adviser is then a source of social influence of the kind Lorenz et al. studied, broadcast to every member at once, and a reliance study is an advice-taking study of the kind Soll and Larrick ran. Collective-judgement benchmarks have been applied to reliance data only in a few probes (He et al., 2025; Fugener et al., 2021; Berger et al., 2026), and none of them reports an accuracy-adjusted error correlation.

The ensemble-learning literature offers the formal bridge. Wood et al. (2023) prove that diversity is a third term in the bias-variance decomposition of an ensemble's loss for a range of losses, measured as a dependence between members, and argue that it should be managed as a trade-off rather than maximised. For the zero-one loss they show that the effect of diversity depends on the label distribution. The quantity Section 3 derives is the human-population analogue of that dependence term. Steyvers et al. (2022) carry the ensemble view into human-machine pairs, using a Bayesian framework to study which factors govern the performance of hybrid combinations of human and machine classifiers, given the different ways each expresses confidence. Their combinations pool independent human and machine judgements; the question here is what happens when the human judgements are no longer independent of the machine, or of each other.

2.4 Why the three have not met

The obstacle is not that nobody has noticed the question. He et al. (2025) state it in their title; Fugener et al. (2021) built a study around it; Schoenegger et al. (2024a) pre-registered a hypothesis on it. The obstacle is that the reliance literature's standard outputs are pooled over cases, the monoculture literature's models treat reliance as binary adoption, and the collective-judgement literature's measures have mostly been applied to independent human judges; the few applications to judges who share an adviser use raw or benchmark-difference measures rather than a correlation. The next section writes down the quantity all three would need to agree on.

3. A minimal model: from reliance rates to error correlation

3.1 Setting and two assumptions

Consider a population of decision-makers who each decide the same set of cases, labelled i , with an AI recommendation shown on each. Write p_i for the probability that a decision-maker drawn from the population errs on case i when assisted, and $p_{0,i}$ for the same probability when unassisted. Write $a_i = 1$ if the AI's recommendation on case i is wrong and 0 if it is right, and e for the fraction of cases on which the AI is

wrong.

The model rests on two assumptions, stated so that they can be attacked.

- **A1 (exchangeability).** Within the population being studied, the decision-makers are interchangeable: the probability of erring on case i does not depend on which person is drawn. Heterogeneous populations can be split into strata (for example by experience), and the analysis applied within each.
- **A2 (conditional independence).** Given the case, two different people's errors are independent. All dependence between people runs through properties of the case: its difficulty, its ambiguity, and the advice shown on it. A2 excludes people who see each other's answers. It does not exclude a shared adviser, because the adviser's output is a property of the case.

A2 is the assumption that makes the bridge possible, and it is also the assumption that a real committee breaks, since members talk. Section 6 returns to that.

3.2 The identity

Under A1 and A2, for two people j and k drawn from the population, the probability that both err on a random case is the mean over cases of p_i squared, and the probability that either errs is the mean of p_i , written m . The covariance between their error indicators is therefore

$$\text{Cov} = \text{mean}_i(p_i^2) - m^2 = \text{Var}_i(p_i),$$

the variance across cases of the case-level error probability. The pairwise error correlation, ϕ , divides by the Bernoulli variance:

$$\phi = \text{Var}_i(p_i) / (m(1 - m)).$$

This is a standard identity in the analysis of binary ratings, not a contribution of this paper. Its consequence for reliance research is the point. Error correlation across people is not a property of any one person. It is a property of how the probability of error is spread over cases: it is zero when every case is equally hard and approaches one when every case is either always right or always wrong. Any metric that is computed per person and then pooled over cases has already averaged away the quantity that determines population-level correlation.

3.3 The decomposition by AI correctness

Split the cases by whether the AI is right. Let f be the mean error probability on the cases where the AI is wrong, and u the mean error probability on the cases where the AI is right. Then $m = e f + (1 - e) u$, and the law of total variance gives

$$\begin{aligned} \text{Var}_i(p_i) = & e(1 - e)(f - u)^2 \\ & \text{(between-class term)} \\ & + e \text{Var}(p_i \mid \text{AI wrong}) \\ & \text{(within-class, AI wrong)} \\ & + (1 - e) \text{Var}(p_i \mid \text{AI right}). \end{aligned}$$

(within-class, AI right)

Each term has a reading in the language of reliance research.

The **between-class term** is the part of the shared error that comes purely from the AI concentrating errors on its own mistakes. On a binary task, f is the over-reliance rate in the strict sense, the rate of accepting wrong advice, and u is the rate of erring despite correct advice, which on a binary task means rejecting it. The term grows with the gap between them and is largest when the AI is wrong on about half of the cases. It is identifiable from three numbers that a reliance study can report: e , f and u .

The **within-class term for AI-wrong cases** measures whether people agree on which wrong suggestions to follow. If every wrong suggestion is equally persuasive, this term is zero. If some wrong suggestions are followed by almost everyone and others by almost no one, it is large. A reliance study that reports only the pooled over-reliance rate cannot recover it.

The **within-class term for AI-right cases** measures whether people agree on which correct suggestions to reject, and also carries whatever shared difficulty survives assistance on cases where the advice was right.

All three terms are non-negative. The between-class term is therefore a lower bound on the assisted error covariance under A1 and A2, and dividing it by $m(1 - m)$ gives a lower bound on ϕ . That bound is computable from published reliance rates. The gap between it and the true value is what item-level records would add.

3.4 The comparison that matters is with the unassisted baseline

A lower bound on assisted correlation does not say whether assistance raised correlation, because humans unaided already share blind spots. He et al. (2025) note that clinicians share biases from common training, and their control groups show positive error dependence. Hedden and Raghavan (2026) make the same point against the monoculture critics: the polyculture baseline is not independent either, since humans make correlated errors through shared culture, systematic biases and overlapping information. The change in covariance due to assistance is

$$\Delta = \text{Var}_i(p_i) - \text{Var}_i(p_{0_i}),$$

and its sign depends on how the AI reshapes the difficulty profile. Two forces act in opposite directions. Assistance flattens the profile on cases that humans find hard and the AI gets right, because those cases stop being shared failures. Assistance sharpens the profile on cases the AI gets wrong, because a persuasive wrong suggestion turns a case of moderate difficulty into a shared failure. Which force dominates depends on the overlap between the humans' hard cases and the AI's wrong cases, a quantity that the error-consistency literature measures between systems (Geirhos et al., 2020) and that Groh et al. (2022) describe qualitatively

for deepfake detection, where people and the leading model were similarly accurate but made different kinds of mistakes. When human and AI errors overlap little, the flattening force is available; when they overlap a great deal, the AI cannot remove the humans' shared blind spots and can only add its own.

This is the reconciliation the paper proposes. The mechanical argument for rising correlation is correct about the between-class term and silent about the flattening force. The deflationary argument is correct that the baseline is not independent and silent about how large the between-class term can become. Which is right in a deployment is an empirical question about e , f , u , the within-class terms and the human-AI overlap.

3.5 Two complications

Multi-class tasks. On a binary task two people who both err give the same answer. On a task with many options they may err differently, and a majority vote is only damaged when they err the same way. Kim et al. (2025) therefore measure the agreement rate given that both models are wrong, and report it against the chance rate implied by the number of wrong options. The identity above is about the co-occurrence of errors, which is what matters for second opinions that check whether a first answer is wrong; for plurality votes the relevant quantity is co-occurrence of the same wrong answer, which a shared adviser makes more likely, since everyone who follows it errs identically.

Accuracy-adjusted metrics. Raw agreement between two judges rises with their accuracy, because two people who are usually right agree on the cases they get right. Geirhos et al. (2020) introduced error consistency to correct for this, comparing observed agreement with what two independent judges of the same accuracies would show; Goel et al. (2025) extend the correction to probabilistic outputs and count different wrong answers as disagreement. Any claim that assistance changed correlation has to be made on an accuracy-adjusted scale, because assistance changes accuracy. Section 5.1 shows that the probes mostly were not.

4. What the direct probes report

Table 1 consolidates the published results that bear most directly on whether AI assistance changes the correlation of human errors, or of human outputs, with the population-level measure each study used and the direction it reported.

Figure 1 plots the two studies whose numbers allow the decomposition of Section 3 to be applied.

4.1 He et al.: assistance in practice reduced shared errors

He et al. (2025) ran two pre-registered field experiments at a tertiary teaching hospital with second-year medical master's students diagnosing lung nodules from CT scans. Thirty cases, fifteen malignant and fifteen benign by histopathology or three-year follow-up, were curated to span three difficulty levels. The AI system was right on

Table 1. Published probes of whether AI assistance correlates human errors or outputs. "Direction" is the direction the source reports for its own population-level measure; "n/a" marks studies that report no population-level measure.

Source	Task, population	AI condition	Population measure	Reported result	Direction
He et al. (2025), Study 1	Lung CT, novices, 84 per arm	AI in practice (A2)	Shared error index; triad penalty d	Index 47.15 to 25.70; d 0.06 to 0.02	Less shared
He et al. (2025), Study 1	Lung CT, novices	AI in training only (B1)	Shared error index; triad penalty d	Index fell (significant); d 0.06 to 0.07 at higher accuracy	Index: less; d: more
He et al. (2025), Study 2	Lung CT, novices, unaided cases	AI early in practice (G2)	Shared error index; d	Index 30.48 to 20.28; d 0.05 to 0.03	Less shared
Schoenegger et al. (2024a)	Forecasting, N = 991	LLM assistant	Aggregate crowd accuracy	Pre-registered: noisy assistant improved the aggregate, superforecasting assistant no different; exploratory (one outlier question removed): noisy assistant reduced aggregate accuracy	No clear change
Fugener et al. (2021)	Experiments plus simulation	AI advice	Simulated crowd accuracy	Assisted crowds far less effective	More shared
Dratsch et al. (2023)	Mammography, 27 readers	12 of 40 suggestions wrong	Accuracy by AI correctness	79.7-82.3% vs 19.8-45.5%	n/a (see 4.4)
Ahmadzade et al. (2025)	Mammography, 14 readers	AI assistance	Kappa vs gold standard	Density kappa 50.01% to 81.38%	n/a (accuracy)
Doshi & Hauser (2024)	Short stories, online	LLM story ideas	Similarity between stories	More similar with AI	More homogeneous
Padmakumar & He (2024)	Argumentative essays	InstructGPT, GPT-3	Lexical, content diversity	Lower with InstructGPT only	More homogeneous
Dong & Yakura (2026)	Metaphors, L1 and L2 writers	AI ideation vs refinement	Collective diversity	Ideation compressed; refinement preserved	Depends on design
Burtch (2026)	Kaggle code, 2019-2026	Field diffusion of AI	Syntax vs approach similarity	Syntax converged; approach did not	Mixed

83.33% of the 30 cases at a 0.5 threshold, so wrong on five. In Study 1, a 2 by 2 design with 84 participants per arm varied AI input in training and in practice. In Study 2, one group had AI input on the first 15 practice cases only, and one had reduced AI input (probabilities without explanatory features) in training only; group-level metrics were computed on the last 15 cases, which every group diagnosed unaided.

The authors report two population-level measures. The first, a shared error index, counts for each error a participant makes how many others in the same arm also erred on that case, averaged over the participant's errors. In Study 1 it fell from 47.15 in the control to 25.70 with AI in practice and 21.24 with AI in both phases; in Study 2 it fell from 30.48 to 20.28 for the group with early AI input. The second is a dependence penalty d: synthetic three-person committees are formed by sampling participants within an arm, a majority vote is taken case by case, and the committee's accuracy is subtracted from what three people of the same individual accuracies would achieve if their errors were independent. Table 2 of the paper reports d of 0.06 for the Study 1 control, 0.02 for AI in practice (A2) and in both phases (B2), and 0.07 for AI in training only (B1); in Study 2, 0.05 for the control (Table 2; the text gives 0.06 for both controls), 0.03 for early AI input in practice (G2) and 0.06 for reduced AI input in training (G3). Figure 1a plots these against individual accuracy.

Two findings matter for this paper. Assisted decisions in Study 1 (A2, B2) were both more accurate and less

dependent than unassisted ones. And not every intervention that raised accuracy lowered dependence: training-only AI input (B1 in Study 1, G3 in Study 2) raised individual accuracy slightly, from 0.68 to 0.71 and from 0.78 to 0.82, while d rose slightly, from 0.06 to 0.07 and from 0.05 to 0.06. On the shared error index the same two arms moved the other way: the authors report a significant fall relative to control for both, smaller than in the arms with AI in practice. The two measures therefore disagree for these arms; Section 5.1 explains why the index, which is not accuracy-adjusted, can fall when accuracy rises. The authors read this as evidence that individual gains do not always become collective gains, and suggest that using AI during practice may have helped participants refine idiosyncratic decision rules, raising competence while leaving their residual errors heterogeneous.

The paper also reports how participants used the advice. When the AI was right, the Study 1 group with AI in both phases (B2) diverged from it 16% of the time and the group with AI in practice only (A2) 23% of the time; when the AI was wrong, the practice-only group diverged from it 67% of the time and the both-phases group 65%, with no significant difference between groups. On this binary task, diverging from wrong advice is a correct answer and diverging from right advice an error, so these figures give f and u directly: f of about 0.33 for A2 and 0.35 for B2, u of 0.23 for A2 and 0.16 for B2.

Worked arithmetic (by the author, from the reported rates). With $e = 5/30$, the decomposition of Section 3.3 predicts a mean error of $e f + (1 - e) u$. For the AI-in-practice group this is 0.247, an accuracy of 0.753 against the reported 0.75; for the both-phases group it is 0.192, an accuracy of 0.808 against the reported 0.81. The reported rates are internally consistent with the model's accounting. The between-class term $e(1 - e)(f - u)^2$ is 0.0014 for the first group and 0.0050 for the second, which divided by $m(1 - m)$ gives lower bounds on the assisted pairwise error correlation of about 0.008 and 0.032. The identifiable part of the AI-induced correlation is small in this study, because people rejected wrong advice about two times in three and the AI was wrong on only one case in six. Whatever residual dependence the assisted groups show must come from the within-class terms: agreement about which of the five wrong suggestions to follow, and shared difficulty that survived correct advice. The paper does not report the case-level rates that would separate them.

4.2 Schoenegger et al.: no detectable damage to the crowd, on six questions

Schoenegger et al. (2024a) assigned 991 participants to a superforecasting-prompted assistant, a deliberately overconfident and base-rate-neglecting assistant, or a weaker control model that gave no numerical forecasts, across six forecasting questions. Their pre-registered analysis found that both frontier assistants improved individual accuracy, by between 24% and 28% relative to control. They pre-registered a hypothesis that assistance would degrade the wisdom of the crowd by reducing the diversity of forecasts, and report that their data do not convincingly show that LLM augmentation reduces, or increases, wisdom-of-the-crowd effects uniformly; pre-registered and exploratory analyses pointed in opposite directions, driven by an outlier question.

The scope of this negative result is narrow, as the authors' own agnosticism about aggregate effects signals. Six questions is a small sample of items for an aggregate-accuracy comparison, and aggregate crowd error conflates the correlation of individual errors with their mean: if assistance lowers everyone's error substantially, the crowd can improve even as dependence rises. The study is evidence that, in this design, the net effect on the crowd was not clearly negative. It is not a measurement of error correlation.

4.3 Fugener et al.: convergence and a simulated crowd penalty

Fugener et al. (2021), in a multi-method study combining an analytical model, experiments and a simulation, report that human choices under AI advice converged toward similar responses and improved individual accuracy, while each person's unique knowledge, what they knew that the AI did not, shrank. Simulations based on their experimental data indicate that groups of humans interacting with the AI were far less effective as crowds than groups without it, and

personalised advice mitigated the loss. This is the clearest published evidence for the mechanical argument. Two features limit its reach for the present question. The crowd results are simulated from individual responses rather than observed in groups, as are He et al.'s committees; and the published summary does not report the AI's error rate or the f and u rates that would locate the study on Figure 1b. The study shows that convergence happens under some designs. It does not show the decomposition of the resulting dependence.

4.4 Dratsch et al.: planted errors create shared errors

Dratsch et al. (2023) asked 27 radiologists, grouped by experience, to assign BI-RADS categories to 50 mammograms with the help of a purported AI system. After ten training cases with correct suggestions, 12 of the 40 test cases carried an incorrect suggestion. Accuracy on cases with correct suggestions was 79.7% for inexperienced, 81.3% for moderately experienced and 82.3% for very experienced readers; on cases with incorrect suggestions it fell to 19.8%, 24.8% and 45.5%. Inexperienced readers were more likely than both experienced groups to follow an incorrect suggestion that overstated the category.

The study reports no population-level measure, but its numbers are exactly the inputs the decomposition needs.

Worked arithmetic (by the author, from the reported accuracies). With $e = 12/40$, $f = 1$ minus accuracy on wrong-suggestion cases and $u = 1$ minus accuracy on correct-suggestion cases, the between-class term divided by $m(1 - m)$ gives lower bounds on the pairwise error correlation within each experience group of about 0.32 (inexperienced), 0.29 (moderately experienced) and 0.14 (very experienced). These are bounds on error co-occurrence; because BI-RADS has several categories, two readers who both err need not choose the same category, although readers who follow the suggestion do. Two cautions apply. The wrong suggestions were planted, so $e = 0.3$ reflects the design and not a deployed system's error rate, and the cases chosen to carry them may not resemble the cases a real model gets wrong. And there is no unaided arm in which to measure the baseline correlation, so the study cannot say whether assistance raised it. What it does show is that a regime with large $f - u$ exists among practising radiologists, and that in that regime the identifiable part of shared error is roughly 4 to 40 times larger than in He et al.

4.5 Agreement that rises with accuracy

Ahmadzade et al. (2025) had 14 radiologists read 65 screening mammograms with and without AI assistance in a crossover design, and report that agreement with the gold standard improved, for example with the kappa for BI-RADS breast density among general radiologists rising from 50.01% to 81.38%, and that AI reduced variability across experience levels. Studies of this kind are often cited as evidence that AI makes readers consistent, which is true and is the intended effect. They are not evidence about error correlation. When everyone's

agreement with the truth rises, agreement with each other rises mechanically; what would bear on the question is whether the readers' remaining errors fell on the same cases more or less often than their accuracies imply, which an accuracy-adjusted error-consistency analysis of the same data could report.

4.6 Homogenised output is a different quantity

A substantial body of work finds that generative assistance makes human output more alike. Writers given LLM story ideas produced stories rated more creative, especially among less creative writers, but more similar to one another (Doshi & Hauser, 2024). In a 36-person study, ChatGPT users produced less semantically distinct ideas than users of an alternative support tool (Anderson et al., 2024). Co-writing with InstructGPT, but not with GPT-3, reduced the diversity of argumentative essays, mostly through the model's contributions (Padmakumar & He, 2024). Writing with an assistant configured to argue one side of a question shifted both what participants wrote and their later attitudes (Jakesch et al., 2023). Sourati et al. (2025) review evidence that LLMs reinforce dominant styles and reasoning strategies. In the field, customer-support agents with a generative assistant gained most when they were less experienced, with small declines in quality for the most skilled (Brynjolfsson et al., 2023), which compresses the performance distribution.

Two recent results complicate the homogenisation story. Dong and Yakura (2026) find in a pre-registered experiment that AI ideation compressed collective diversity, while AI refinement of writers' own ideas preserved it, so the design of the workflow decided the outcome. Burtch (2026), studying Kaggle submissions from 2019 to mid-2026, finds that code became more alike in surface syntax while showing no evidence of convergence in approach or intent. Burtch separates a surface measure from a deeper one and finds that the alarming result holds for the surface.

None of these studies measures error. Output similarity is the content analogue of agreement, not of error correlation: two essays can be alike and both correct, and two diagnoses can differ and both be wrong. The homogenisation evidence is relevant to the present question only through a hypothesis that it does not test: that people whose outputs converge on a model's will also inherit the model's errors on the same cases.

4.7 Evidence on how people use advice, case by case

Several studies report the case-dependence of reliance that the within-class terms describe, without computing a population quantity. Reverberi et al. (2022), with 21 endoscopists reviewing 504 colonoscopy videos with and without AI, report that endoscopists followed the AI more when it was correct than when it was wrong, through a weighted integration of their own and the AI's opinions that tracked case-by-case reliability. Groh et al. (2022), with 15,016 participants, found human observers and a leading deepfake detector similarly accurate but making different kinds of mistakes, with

machine-informed participants more accurate than either alone while inaccurate model predictions often lowered participants' accuracy. Yu et al. (2024) found that the occurrence of AI errors was a strong influence on whether assistance helped a radiologist. Stevenson and Doleac (2024) found that judges given an algorithmic risk assessment changed their sentencing but deviated from its recommendations in systematic ways, so that the tool produced no detectable gain in public safety or incarceration. Chiang et al. (2023) found that groups making AI-assisted recidivism judgements relied on the AI more than individuals regardless of its correctness, while being more confident when they overturned wrong recommendations. Each of these is a statement about where, across cases or across people, reliance falls. None is a statement about whether two people's residual errors coincide.

5. Why the probes disagree: four confounds

The probes in Table 1 are often cited against one another, as if one showed that AI assistance correlates errors and another that it does not. Read against Section 3, they mostly measured different things, in different regimes, on scales that are not comparable. Four confounds account for most of the apparent conflict.

5.1 Unadjusted measures move with accuracy

He et al.'s shared error index is a raw count of other participants who erred on the same case. When an arm becomes more accurate, fewer people err on any case, so the count falls even if errors remain exactly as concentrated as before. The fall from 47.15 to 21.24 in Study 1 mixes two effects, higher accuracy and lower concentration, and the index alone cannot separate them.

The dependence penalty d corrects for this by comparing with an independence benchmark at the same accuracies, which is the right move, but it is still not scale-free. Under A1 and A2, for a three-person majority with mean individual error m , the gap between the independence benchmark and the correlated committee's accuracy is exactly

$$d = 3(1 - 2m) \text{Var}_i(p_i) - 2\mu_3,$$

where μ_3 is the third central moment of the case-level error probabilities. Substituting $\text{Var}_i(p_i) = \phi m(1 - m)$, the leading term is $3g(m)\phi$ with $g(m) = (1 - 2m)m(1 - m)$. For a fixed error correlation ϕ , g peaks near $m = 0.21$ and falls to zero both as accuracy approaches one (m to 0) and as it approaches chance on a binary task (m to 0.5), so d is not comparable across accuracy levels in general, and a smaller d is not by itself a smaller correlation. Over the range of He et al.'s Study 1 the scaling cannot produce their result, because it runs the other way: g is about 0.078 at the control's m of 0.32, about 0.094 at A2's 0.25 and about 0.095 at B2's 0.19 (worked arithmetic by the author). Since g is flat or rising over that range while d fell threefold, the fall is not an artefact of accuracy scaling at this level of

approximation, and the direction of He et al.'s result survives. Its magnitude, in correlation units, is not established by the published numbers, because the third-moment term and the heterogeneity of participants within arms are unknown. The derivation also assumes exchangeable committee members, which synthetic committees drawn from a heterogeneous arm only approximate.

Inter-reader agreement in the mammography crossover study (Ahmadzade et al., 2025) and aggregate crowd accuracy in the forecasting study (Schoenegger et al., 2024a) share the same problem in stronger form. Neither separates the mean of errors from their dependence. The correction is known: error consistency (Geirhos et al., 2020), chance-adjusted agreement on mistakes (Goel et al., 2025), or the variance decomposition above computed on case-level data. Kim et al. (2025) faced the same issue for model-model comparisons and conditioned on both models being wrong for that reason.

5.2 The regime is set by e , f and u , and the probes sit in different regimes

Figure 1b shows the second confound. In He et al. (2025), the accuracy gap between cases with right and wrong advice was about 10 to 19 points; in Dratsch et al. (2023), 37 to 60 points. The AI was wrong on one case in six in the first study and, by design, on three in ten in the second. The between-class term grows with the square of the gap and with $e(1 - e)$, so the two studies differ by roughly 4 to 40 times in the identifiable part of AI-induced correlation, before any within-class term is counted.

Some of this difference is design. Dratsch et al. presented a purported AI and planted wrong suggestions; He et al. used a real detection-and-classification system with explanatory features and probabilities, and their participants had a lecture and, in some arms, AI-supported training. The design differences are the kind of thing that moves f : Reverberi et al. (2022) found endoscopists followed AI more when it was right than when it was wrong, through case-by-case weighting, which keeps f low, whereas automation bias studies collected by Goddard et al. (2012) document conditions such as workload, time pressure and inexperience under which f rises. Both regimes exist among professionals. A general claim that assistance does or does not correlate errors is a claim about where deployments sit on Figure 1b, and the probes say that they can sit in very different places.

5.3 Outputs, errors and approaches are different quantities

The third confound is the conflation of output homogenisation with error correlation, discussed in Section 4.6. Burtch (2026) shows that the measure chosen decides whether homogenisation appears, and Dong and Yakura (2026) show that the workflow design decides it. For decision tasks the relevant measure is neither surface similarity nor approach similarity but co-occurrence of errors on the same cases. A study of essays

cannot supply it, because essays have no ground truth on which to err.

5.4 Synthetic groups are not groups

The fourth confound is that every population-level result for humans in Table 1 is computed from individual decisions recombined after the fact. He et al. (2025) form synthetic committees by sampling participants; Fugener et al. (2021) simulate crowds from experimental data; Berger et al. (2026) evaluate their independence-preserving protocol by generating all pairs of unaided humans per case. Synthetic recombination is exactly what A2 requires, and it is the right tool for estimating the variance term. But it measures the potential of independent pooling, not what happens when assisted people deliberate. Real groups add a channel the model excludes. Chiang et al. (2023) found that groups relied on AI more than individuals regardless of its correctness, which in the terms of Section 3 raises f for the group decision; Lorenz et al. (2011) found that seeing others' estimates narrowed diversity without improving collective error. Whether an AI adviser and a peer discussion compound, cancel or substitute for each other is unmeasured.

A related caution concerns inference. Significance tests over thousands of resampled committees drawn from a fixed set of participants and cases describe the resampling distribution, not the uncertainty about a new population of people and cases. With 30 cases, of which five carried wrong advice, the precision with which any study of this size can estimate a case-level variance is limited by the cases, not the resamples.

6. When does correlation cost anything?

The mechanical argument treats rising correlation as a cost. Whether it is depends on what consumes the independence. Four structures are distinguished here, because they give different answers.

One case, one decider. If each case is decided once, by one assisted person, and nothing downstream compares decisions across people, then error correlation across people has no effect on accuracy. Each decision is right or wrong on its own. Correlation still matters for two things outside accuracy: the variance of total errors in a period, since correlated errors arrive together when the AI's failure mode is hit, and the detectability of errors, since an audit that samples a second opinion from the same assisted population will tend to confirm the first.

Sequential second opinion. If a positive or doubtful first reading triggers a second, as in the lung-cancer workflow He et al. (2025) describe from interviews with 36 physicians, the value of the second reading depends directly on the probability that it errs given that the first erred. That is the conditional error rate the covariance controls. Under A1 and A2, the probability that the second reader errs given that the first did is $m + \text{Var}_i(p_i)/m$. When the AI is shown to both readers, the between-class term enters this conditional probability in full.

Parallel aggregation. Committees, crowds and ensembles of human judges are where the Condorcet logic applies. Ladha (1992) shows that small groups are the sensitive case, which is the clinical case: second and third opinions, tumour boards. Kurvers et al. (2016) show that the gains from pooling also depend on similar accuracy among members, and Kurvers et al. (2023), pooling ten independent diagnosticians per case on 1,333 cases from an online platform, report accuracy rising from 46% for single diagnosticians to 76% for the pooled ten. The value lost if a shared adviser correlated those ten would be large. Berger et al. (2026) take the design route: their hybrid confirmation tree elicits a human and an AI judgement independently, accepts them if they agree and otherwise asks a second human, and across ten datasets with over 41,000 judgements it beat the AI-as-advisor arrangement in every dataset, by 0.2 to 6.6 percentage points and by 4.45 points pooled. Using a signal-detection model, the authors attribute the gain to two shortfalls of the advisor arrangement: people cannot discriminate well enough between correct and incorrect advice, and overall they rely on it too little. In Section 3's terms, the tree keeps the human judgements out of the AI's influence and so removes the between-class term from the human side altogether. Zoller et al. (2024) report a related result for open-ended diagnosis: collectives combining physicians and five LLMs, on 2,133 cases, beat physicians, physician collectives, single LLMs and LLM ensembles, which the authors attribute to humans and models making different kinds of errors.

The person decided about. When one applicant faces many decision-makers, as in hiring or lending, correlation across decision-makers is outcome homogenization in Bommasani et al.'s (2022) sense: the same person rejected everywhere. Here the cost is not to the decision-makers' accuracy but to the applicant, and it is the concern of Creel and Hellman (2022) and of the case for algorithmic pluralism (Jain et al., 2024). Hedden and Raghavan (2026) argue that in markets with capacity constraints, roughly the same number of candidates go unhired either way, so monoculture redistributes rather than adds exclusion; they also note that a single ensemble algorithm can in principle absorb the information a polyculture would aggregate. Those arguments bear on fully algorithmic decisions. With partial human reliance the shared component is not an ensemble that pools many signals; it is one model's output, added on top of human judgements that already share blind spots.

The deflationary objections deserve their full weight in each structure. If firms can choose whether to adopt a common technology, Hedden and Raghavan argue that the welfare loss from monoculture at equilibrium is modest outside pathological cases. Raghavan (2024) shows in a game-theoretic model that competition rewards distinctive output and so pushes toward diverse models. Davis-Stober et al. (2014) find that crowds of biased and correlated judges usually remain wiser than a randomly chosen member. Each of these limits how bad

correlation can be. None says that it is absent, and none applies to the sequential second-opinion structure, where there is no market and no large crowd.

7. A decision procedure

Algorithm 1 states the decision structure the paper analyses: given a deployment of AI advice to many human decision-makers, which quantities must be measured before claiming that assistance is safe or unsafe at the population level.

Two features of the procedure are deliberate. It returns "no verdict" in two places, because a claim about the change in correlation without an unaided baseline on the same cases, or without case-level records, is not supported by any of the quantities that would be needed. And its final line forbids reporting an accuracy gain without the dependence quantities, because Section 5.1 shows that a gain in accuracy moves every unadjusted dependence measure in the reassuring direction.

8. Measurements that would settle the open parts

The measurements below are stated so that each can be run, and so that its outcome would change a specific claim in this paper.

R1. Reanalyse existing aided and unaided datasets on an accuracy-adjusted scale. The ten datasets assembled by Berger et al. (2026) cover dermatology, colonoscopy, review sentiment, review deception, deepfake detection, news truthfulness and rearrest prediction, and include per-case decisions by many people with and without AI advice. Among their sources are Lai and Tan (2019), who varied machine assistance along a spectrum from explanations alone to predicted labels for deception detection; Bansal et al. (2021); Groh et al. (2022); Reverberi et al. (2022); and Fogliato et al. (2021), on rearrest predictions by laypeople given a risk assessment. For each, computing the case-level variance of error under both conditions, the covariance change Δ , and the change Δ_{ϕ} in the accuracy-adjusted correlation ϕ would directly test whether assistance raised or lowered the correlation of human errors across ten tasks. The paper predicts that the sign of Δ_{ϕ} will vary across datasets, and that it will be predicted by the overlap O between the cases unaided humans find hard and the cases the AI gets wrong. A uniform sign of Δ_{ϕ} across all ten would contradict that prediction. The sign of Δ alone cannot test it, because when error rates are below one half a gain in accuracy shrinks $m(1 - m)$ and so lowers the covariance even if ϕ is unchanged.

R2. Decompose the assisted variance. On the same datasets, split assisted variance into the between-class term and the two within-class terms. The size of the within-class term for AI-wrong cases measures whether people agree on which wrong suggestions to follow. If it is large, the persuasiveness of a wrong suggestion is a property of the case and not of the person, and

Algorithm 1. Does AI assistance put population-level accuracy at risk through correlated errors? Input: deployment D (cases, decision-makers, AI, how decisions are combined) Output: a verdict, and the measurement that would change it

```

1. Identify how D consumes independence:
  if each case is decided once and never re-checked by another person:
    return "correlation does not affect accuracy; monitor error bursts and audit design"
  if a second reader is consulted conditionally (second opinion):
    target <- P(second errs | first errs)
  if decisions are pooled (committee, crowd, vote):
    target <- accuracy-adjusted pairwise error correlation phi
  if one person faces many decision-makers (hiring, lending):
    target <- outcome homogenization across decision-makers
2. Check that a human baseline exists: same cases, same population, unaided.
  if not: return "no verdict possible on the change; only assisted levels"
3. From reliance data, estimate e, f = P(err | AI wrong), u = P(err | AI right).
  B <- e(1 - e)(f - u)^2 # between-class term, a lower bound on assisted covariance
4. If case-level records exist (person x case x aided/unaided):
  V_a <- Var over cases of assisted error rate
  V_0 <- Var over cases of unaided error rate
  Delta <- V_a - V_0 # change in pairwise covariance under A1, A2
  phi_a <- V_a / (m_a(1 - m_a)); phi_0 <- V_0 / (m_0(1 - m_0)) # m_a, m_0: mean error rates
  Delta_phi <- phi_a - phi_0 # change in accuracy-adjusted error correlation
  O <- error consistency between AI and unaided humans (overlap of hard cases)
else:
  return "only bound B is known; request case-level records"
5. If members of the deployment interact (deliberate, see each other's answers):
  flag "A2 violated; synthetic-group estimates do not bound the pooling value of interacting groups"
6. Verdict:
  if Delta_phi <= 0: assistance did not raise the error correlation in this population
  if Delta_phi > 0 and target is pooled or sequential:
    consider independence-preserving designs (elicit before advice; AI as tiebreaker)
  if Delta_phi > 0 and target is outcome homogenization:
    report homogenization per applicant; consider pluralistic or independent-review designs
  # Delta (covariance) stays reported: it sets the second-opinion conditional m + Var/m
  report Delta_phi, Delta, B, O, and the accuracy change together; never the accuracy change alone

```

interventions that target people (training, forcing functions) will not reduce it as much as interventions that target the advice on those cases.

R3. Pre-register a population-level outcome in new reliance experiments. A reliance experiment that has several people decide each case already has what is needed. Reporting e, f and u per condition, and the case-level error rates, costs nothing extra in data collection. He et al. (2025) show that this can be done within a pre-registered field experiment.

R4. Compare synthetic and real groups. Run the same cases with real committees that deliberate, with and without a shared AI adviser, alongside synthetic committees drawn from individual arms. The difference measures how much interaction adds to or removes from the dependence that the shared adviser induces, which Section 5.4 identifies as unmeasured.

R5. Track the regime over time. f and u are not fixed. He et al. (2025) find that AI-supported training changed how participants used advice later, and Yu et al. (2024) find that experience does not predict benefit. Measuring e, f, u and Delta repeatedly in a deployment would say whether populations drift toward the high-correlation regime of Dratsch et al. (2023) as familiarity grows.

R6. Extend to open-ended and multi-model settings. For open-ended tasks, measure co-occurrence of the same wrong answer, as Kim et al. (2025) do for models. Where a population uses several models, the shared component is the models' correlated errors, which Kim et al. find rise with capability; Schoenegger et al. (2024b) report that the aggregate of twelve LLMs on 31 binary

questions was not statistically different from a crowd of 925 human forecasters, which makes the question of how human and model crowds combine practical.

9. Objections

"The identity assumes away the interesting part." A1 and A2 do exclude real interaction and person-level heterogeneity. They are the assumptions under which synthetic-committee analyses, including those of He et al. (2025) and Fugener et al. (2021), are already valid; the paper makes them explicit. Heterogeneity can be handled by stratifying, and Section 5.4 states what interaction adds that the model cannot.

"The worked numbers are not measurements." They are not, and the paper does not present them as such. They are arithmetic on published rates, under stated assumptions, and they are a lower bound. Their role is to show that two published studies sit roughly 4 to 40 times apart on the identifiable part, which is enough to explain why the probes disagree without any claim about the within-class terms.

"Higher accuracy with higher correlation is still better." Often it is. A population that is right 90% of the time with correlated errors may beat one that is right 70% of the time with independent errors, even as a committee. The paper does not claim otherwise. The claim is that reliance research cannot currently say which trade it is making, because it reports one side of it. Wood et al. (2023) make the same argument for ensembles: diversity is one term in a trade-off, to be managed, not maximised.

"Monoculture critics were already refuted." Hedden and Raghavan (2026) show that several objections to monoculture fail and that others are not decisive in general. Their analysis concerns fully algorithmic decisions; they themselves note that the polyculture baseline is correlated. This paper's decomposition is consistent with their point about the baseline and adds the case they do not treat: partial reliance by humans whose shared blind spots the AI may remove or add to.

"This is just automation bias." Automation bias is a property of a person: their tendency to follow wrong advice. The between-class term is built from it. The population-level effect needs one more ingredient that automation-bias studies do not measure: whether different people are fooled by the same suggestions.

10. Related work and how this paper differs

The nearest formal work is Tan (2026), which models each employee's judgement as a weighted combination of an idiosyncratic estimate and AI advice, derives team-level error variance and an effective number of independent judgements, and illustrates a paradox in which individually optimal reliance can raise team error. That model takes reliance as a weight applied uniformly to every case. The present paper's starting point is the empirical fact that reliance is case-dependent, and its contribution is to show that the population quantity lives in that case-dependence, to identify which part of it published reliance rates can recover, and to apply that to published data. Biswas (2026) models reliance as a population process in which agents learn about AI quality from experience and from peers, and finds that visible unverified use can tip a population into collective over-reliance; the outcome there is the population rate of over-reliance, not the correlation of errors across people on the same cases.

Fugener et al. (2021) and He et al. (2025) are the two empirical papers closest to this one, and they reach opposite conclusions. This paper does not adjudicate between them by preferring one; it locates them in a common model and shows what each did not report. Hedden and Raghavan (2026) and Raghavan (2024) supply the deflationary side for algorithmic decision-making; Bommasani et al. (2022), Toups et al. (2023) and Kim et al. (2025) supply the measurement tradition for models. Sourati et al. (2025) review homogenisation of expression and thought, which this paper distinguishes from error correlation. Berger et al. (2026) are closest in spirit to the remedy, since their protocol removes the shared adviser from the human side, and their compiled datasets are the most direct route to running R1.

Within this author's own work, earlier papers examined correlated failure between AI checkers and the AI they check, shared memory among LLM agents, and the effect of explanations on human over-reliance. This paper is about a different population and a different quantity: many humans sharing one adviser, and the correlation of their residual errors.

11. What is not known

No published study reports, for human decision-makers on the same cases, a scale-free (correlation-scale) accuracy-adjusted error measure with and without AI assistance. The nearest existing quantity is He et al.'s (2025) dependence penalty d , which is benchmarked against independence at the same accuracies but, as Section 5.1 shows, is not comparable across accuracy levels. The sign of the change is therefore not known for any deployment. The within-class terms, how strongly people agree on which wrong suggestions to follow and which right ones to reject, have not been reported for any reliance dataset. The overlap between the cases unaided humans find hard and the cases a deployed AI gets wrong has not been reported alongside reliance rates. Nothing published compares the dependence induced by a shared AI adviser in deliberating groups with the dependence in synthetic groups drawn from the same people. Whether populations move toward or away from the high-correlation regime with longer use has not been measured.

12. Conclusion

Reliance research asks whether a person uses AI advice well. Monoculture research asks whether a population of decision-makers fails together. The two have not been joined because the first reports quantities pooled over cases and the second models adoption as all or nothing. Under two explicit assumptions, the joining quantity is the variance across cases of the probability of error. Its between-class part is computable from three numbers reliance studies can report, and on published data it differs by roughly 4 to 40 times between a study where wrong advice was usually rejected and one where most readers erred on the cases that carried it. Its within-class parts, and the comparison with an unaided baseline on the same cases, are not in the literature. Those are the measurements that would turn the question from a dispute between an alarm and a deflation into an estimate, and the data to make several of them already exist.

References

- Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology. NBER Working Paper 31422. doi:10.3386/w31422.
- Ahmadzade, M., Rouientan, H., Abdi, N., Norouzi, M., et al. (2025). AI assistance improves reader agreement in digital mammography: A multireader crossover study of general and breast subspecialty radiologists. *Scientific Reports*, 16, 1370. doi:10.1038/s41598-025-31065-3.
- Anderson, B. R., Shah, J. H., & Kreminski, M. (2024). Homogenization Effects of Large Language Models on Human Creative Ideation. *Creativity and Cognition* 2024. arXiv:2402.01536. doi:10.48550/arXiv.2402.01536.
- Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. CHI 2021. arXiv:2006.14779. doi:10.48550/arXiv.2006.14779.

- Berger, J., Analytis, P. P., Satopaa, V., & Kurvers, R. H. J. M. (2026). Beyond AI advice – independent aggregation boosts human-AI accuracy. arXiv:2603.29866. doi:10.48550/arXiv.2603.29866.
- Biswas, A. (2026). Modeling AI Overreliance as a Complex Adaptive System. arXiv:2608.19616. doi:10.48550/arXiv.2608.19616.
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv:2108.07258. doi:10.48550/arXiv.2108.07258.
- Bommasani, R., Creel, K. A., Kumar, A., Jurafsky, D., & Liang, P. (2022). Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization? NeurIPS 2022. arXiv:2211.13972. doi:10.48550/arXiv.2211.13972.
- Broomell, S. B., & Budescu, D. V. (2009). Why Are Experts Correlated? Decomposing Correlations Between Judges. *Psychometrika*, 74(3), 531-553. doi:10.1007/s11336-009-9118-z.
- Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative AI at Work. arXiv:2304.11771. doi:10.48550/arXiv.2304.11771.
- Bucinka, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. arXiv:2102.09692. doi:10.48550/arXiv.2102.09692.
- Burtch, G. (2026). The Hitchhiker's Guide to Monoculture: AI Homogenizes Syntax, Not (Necessarily) Semantics. arXiv:2607.13077. doi:10.48550/arXiv.2607.13077.
- Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2023). Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. CHI 2023. doi:10.1145/3544548.3581015.
- Creel, K., & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, 52(1), 26-43. doi:10.1017/can.2022.3.
- Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, 1(2), 79-101. doi:10.1037/dec0000004.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114-126. doi:10.1037/xge0000033.
- Dong, M., & Yakura, H. (2026). Human diversity fuels collective creativity that large language models cannot simulate or sustain. arXiv:2607.26899. doi:10.48550/arXiv.2607.26899.
- Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28), eadn5290. doi:10.1126/sciadv.adn5290.
- Dratsch, T., Chen, X., Rezade Mehrizi, M., Kloeckner, R., Maehringer-Kunz, A., Puesken, 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. doi:10.1148/radiol.222176.
- Fogliato, R., Chouldechova, A., & Lipton, Z. (2021). The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2). arXiv:2109.01443. doi:10.48550/arXiv.2109.01443.
- Fugener, A., Grahl, J., Gupta, A., & Ketter, W. (2021). Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly*, 45(3), 1527-1556. doi:10.25300/MISQ/2021/16553.
- Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *NeurIPS 2020*. arXiv:2006.16736. doi:10.48550/arXiv.2006.16736.
- Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1), 121-127. doi:10.1136/amiajnl-2011-000089.
- Goel, S., Struber, J., Auzina, I. A., Chandra, K. K., Kumaraguru, P., Kiela, D., Prabhu, A., Bethge, M., & Geiping, J. (2025). Great Models Think Alike and this Undermines AI Oversight. arXiv:2502.04313. doi:10.48550/arXiv.2502.04313.
- Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake Detection by Human Crowds, Machines, and Machine-informed Crowds. *Proceedings of the National Academy of Sciences*, 119(1). arXiv:2105.06496. doi:10.48550/arXiv.2105.06496.
- Guo, Z., Wu, Y., Hartline, J., & Hullman, J. (2024). A Decision Theoretic Framework for Measuring AI Reliance. arXiv:2401.15356. doi:10.48550/arXiv.2401.15356.
- He, V. F., Li, S., Puranam, P., & Lin, F. (2025). Predictive AI Can Support Human Learning while Preserving Error Diversity. arXiv:2502.16411. doi:10.48550/arXiv.2502.16411.
- Hedden, B., & Raghavan, M. (2026). Algorithmic Monoculture and its Critics. arXiv:2604.06047. doi:10.48550/arXiv.2604.06047.
- Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46), 16385-16389. doi:10.1073/pnas.0403723101.
- Jain, S., Suriyakumar, V., Creel, K., & Wilson, A. (2024). Algorithmic Pluralism: A Structural Approach To Equal Opportunity. *FACCT 2024*. arXiv:2305.08157. doi:10.48550/arXiv.2305.08157.
- Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-Writing with Opinionated Language Models Affects Users' Views. *CHI 2023*. arXiv:2302.00560. doi:10.48550/arXiv.2302.00560.
- Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated Errors in Large Language Models. *ICML 2025*. arXiv:2506.07962. doi:10.48550/arXiv.2506.07962.
- Kleinberg, J., & Raghavan, M. (2021). Algorithmic Monoculture and Social Welfare. *Proceedings of the National Academy of Sciences*, 118(22). arXiv:2101.05853. doi:10.48550/arXiv.2101.05853.
- Kurvers, R. H. J. M., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., Argenziano, G., Zalaudek, I., & Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, 113(31), 8777-8782. doi:10.1073/pnas.1601827113.
- Kurvers, R. H. J. M., Nuzzolese, A. G., Russo, A., Barabucci, G., Herzog, S. M., & Trianni, V. (2023). Automating hybrid collective intelligence in open-ended medical diagnostics. *Proceedings of the National Academy of Sciences*, 120(34), e2221473120. doi:10.1073/pnas.2221473120.
- Ladha, K. K. (1992). The Condorcet Jury Theorem, Free Speech, and Correlated Votes. *American Journal of Political Science*, 36(3). doi:10.2307/2111584.
- Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., & Tan, C. (2021). Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. arXiv:2112.11471. doi:10.48550/arXiv.2112.11471.
- Lai, V., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. *FAT* 2019*. arXiv:1811.07901. doi:10.48550/arXiv.1811.07901.
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46(1), 50-80. doi:10.1518/hfes.46.1.50.30392.
- Lorenz, J., Rauhut, H., Schweitzer, F., & Helbing, D. (2011). How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22), 9020-9025. doi:10.1073/pnas.1008636108.
- Marx, C. T., Calmon, F. du P., & Ustun, B. (2020). Predictive Multiplicity in Classification. *ICML 2020*. arXiv:1909.06677. doi:10.48550/arXiv.1909.06677.
- Padmakumar, V., & He, H. (2024). Does Writing with Language Models Reduce Content Diversity? *ICLR 2024*. arXiv:2309.05196.

- doi:10.48550/arXiv.2309.05196.
- Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, 39(2), 230-253. doi:10.1518/001872097778543886.
- Peng, K., & Garg, N. (2024). Monoculture in Matching Markets. *NeurIPS 2024*. arXiv:2312.09841. doi:10.48550/arXiv.2312.09841.
- Raghavan, M. (2024). Competition and Diversity in Generative AI. arXiv:2412.08610. doi:10.48550/arXiv.2412.08610.
- Reverberi, C., Rigon, T., Solari, A., Hassan, C., Cherubini, P., et al. (2022). Experimental evidence of effective human-AI collaboration in medical decision-making. *Scientific Reports*, 12, 14952. doi:10.1038/s41598-022-18751-2.
- Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *IUI 2023*. arXiv:2302.02187. doi:10.48550/arXiv.2302.02187.
- Schoenegger, P., Park, P. S., Karger, E., Trott, S., & Tetlock, P. E. (2024a). AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. arXiv:2402.07862. doi:10.48550/arXiv.2402.07862.
- Schoenegger, P., Tuminauskaite, I., Park, P. S., & Tetlock, P. E. (2024b). Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy. arXiv:2402.19379. doi:10.48550/arXiv.2402.19379.
- Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 780-805. doi:10.1037/a0015145.
- Sourati, Z., Ziabari, A. S., & Dehghani, M. (2025). The Homogenizing Effect of Large Language Models on Human Expression and Thought. arXiv:2508.01491. doi:10.48550/arXiv.2508.01491.
- Stevenson, M. T., & Doleac, J. L. (2024). Algorithmic Risk Assessment in the Hands of Humans. *American Economic Journal: Economic Policy*, 16(4), 382-414. doi:10.1257/pol.20220620.
- Steyvers, M., & Kumar, A. (2023). Three Challenges for AI-Assisted Decision-Making. *Perspectives on Psychological Science*, 19(5), 722-734. doi:10.1177/17456916231181102.
- Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human-AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11), e2111547119. doi:10.1073/pnas.2111547119.
- Tan, K. H. (2026). Cognitive Monoculture and the Individual-Team Reliance Paradox: A Formal Model of Shared AI Dependence in Organizational Decision Making. *International Journal of Web of Multidisciplinary Studies*, 3(8), 465-480. doi:10.71366/ijwos03082685268.
- Toups, C., Bommasani, R., Creel, K. A., Bana, S. H., Jurafsky, D., & Liang, P. (2023). Ecosystem-level Analysis of Deployed Machine Learning Reveals Homogeneous Outcomes. *NeurIPS 2023*. arXiv:2307.05862. doi:10.48550/arXiv.2307.05862.
- Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*. arXiv:2405.06087. doi:10.48550/arXiv.2405.06087.
- Vasconcelos, H., Joerke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M., & Krishna, R. (2023). Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *CSCW 2023*. arXiv:2212.06823. doi:10.48550/arXiv.2212.06823.
- Vodrahalli, K., Daneshjou, R., Gerstenberg, T., & Zou, J. (2022). Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. *AIES 2022*. arXiv:2107.07015. doi:10.48550/arXiv.2107.07015.
- Wood, D., Mu, T., Webb, A., Reeve, H., Lujan, M., & Brown, G. (2023). A Unified Theory of Diversity in Ensemble Learning. *Journal of Machine Learning Research*, 24(359). arXiv:2301.03962. doi:10.48550/arXiv.2301.03962.
- Wu, F., Black, E., & Chandrasekaran, V. (2024). Generative Monoculture in Large Language Models. arXiv:2407.02209. doi:10.48550/arXiv.2407.02209.
- Yu, F., Moehring, A., Banerjee, O., Salz, T., Agarwal, N., & Rajpurkar, P. (2024). Heterogeneity and predictors of the effects of AI assistance on radiologists. *Nature Medicine*, 30(3), 837-849. doi:10.1038/s41591-024-02850-w.
- Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *FAT* 2020*. arXiv:2001.02114. doi:10.48550/arXiv.2001.02114.
- Zoller, N., Berger, J., Lin, I., Fu, N., Komarneni, J., Barabucci, G., Laskowski, K., Shia, V., Harack, B., Chu, E. A., Trianni, V., & Kurvers, R. H. J. M. (2024). Human-AI collectives produce the most accurate differential diagnoses. arXiv:2406.14981. doi:10.48550/arXiv.2406.14981.