Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human Judgment in the Age of Algorithms: Decision Autonomy, Digital Oversight and Organizational Performance

Sarfaraz Abu Abdullah

rajon3430@gmail.com

University of Rajshahi https://orcid.org/0009-0000-3826-1626

Research Article

Keywords: Human Judgment, Algorithmic Decision-Making, Human Oversight, Automation Bias, Human–AI Collaboration, Organizational Performance, AI Governance

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11811714/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Human Judgment in the Age of Algorithms: Decision Autonomy, Digital Oversight and Organizational Performance

[Sarfaraz Abu Abdullah]
[Department of Management, Rajshahi University, Rajshahi, Bangladesh] · Email: [rajon3430@gmail.com]

Abstract

Organizations that put algorithms into consequential decisions usually keep a person in the loop, on the assumption that the person adds something the model lacks. This paper checks that assumption against four bodies of published evidence: a meta-analysis of 106 experiments, field and trial evidence on judges, radiologists, physicians and consultants, a 2025 global survey of 1,993 respondents, and the oversight provisions of the European Union's Artificial Intelligence Act. Published effects are put on a common scale and a short decision model is derived to show when an override can pay. On average a human–AI combination does worse than the better of its two members (Hedges' $g = -0.23$, 95% CI −0.39 to −0.07), which corresponds to a 43.5% chance of beating it. The loss is concentrated in decision tasks ($g = -0.27$); in creation tasks the estimate is positive but imprecise ($g = 0.19$, CI −0.09 to 0.48). Relative capability matters most: combinations gain when the human alone is the stronger member ($g = 0.46$) and lose when the algorithm is ($g = -0.54$). In the model, unrestricted overrides change accuracy by exactly the accuracy gap between human and algorithm, and with unequal error costs the human's odds of being right must exceed the algorithm's by the cost ratio. In the survey, respondents at high-performing organizations were 2.8 times as likely to report a defined human-validation process (65% against 23%; odds ratio 6.2), an association that is large but cross-sectional and self-reported. The evidence is highly heterogeneous ($I^2 = 97.7\%$) and mostly predates current generative models. Oversight helps where it is built around what the human knows and the model does not, and not where it exists in name only.

Keywords

Human Judgment, Algorithmic Decision-Making, Human Oversight, Automation Bias, Human–AI Collaboration, Organizational Performance, AI Governance

1. Introduction

Most organizations that put an algorithm into a consequential decision also promise that a person will stay in the loop. A loan officer reviews the credit score. A judge weighs a risk assessment. A radiologist reads the image next to the model's prediction, and a manager approves the schedule the software proposes. The promise has two justifications. The practical one is that people know things models do not: context, exceptions, the real cost of an unusual error. The legal and moral one is that decisions affecting people should rest with someone who can be held to account, a principle written into data protection law and into the human oversight provisions of the European Union's Artificial Intelligence Act.

Both justifications rest on an empirical premise that is seldom tested: that a person working with an algorithm decides better than either would alone. The older evidence is not encouraging. Simple statistical rules matched or beat expert judgment in many prediction tasks (Meehl, 1954; Dawes et al., 1989), and the newer experimental literature on human–AI teams finds that the combination usually falls short of its better member (Vaccaro et al., 2024). Organizations nonetheless invest in oversight as though it were free and always effective.

1.1. Statement of the Problem

The problem is a managerial one. Leaders who deploy decision algorithms need to know when the human step improves outcomes, when it only adds noise, and how to tell the two apart. They also need to know whether oversight practices are related to the organizational results they are expected to deliver. Published evidence bears on each of these questions, but it sits in separate literatures: behavioral studies of reliance, laboratory experiments, field studies of expert discretion, practitioner surveys and legal analysis. It has not been placed on a common scale, and the conditions under which a human override pays have rarely been stated in a form that an organization can check against its own data.

1.2. Research Questions and Purpose

The paper asks five questions.

RQ1 Do human–AI combinations outperform their best member?

RQ2 What moderates the answer, in particular task type and the relative capability of the human and the algorithm?

RQ3 Under what conditions does a human override improve accuracy?

RQ4 What do field and trial evidence show about expert discretion over algorithmic recommendations?

RQ5 How are human-validation processes related to organizational performance, and what does the evidence imply for regulatory oversight requirements?

The purpose is to assemble published evidence at four levels (experiments, field and trial studies, an organizational survey and the regulatory framework), to express it in comparable metrics, and to derive a rule that organizations can use to judge whether oversight is adding value.

1.3. Scope and Contribution

The analysis works with published results and collects no new primary data. It does not estimate the causal effect of oversight on performance, and the experimental tasks differ from organizational decisions in stakes, accountability and time. Within those limits the paper adds three things. First, it re-expresses the meta-analytic effects with the test statistics and heterogeneity the original authors report, and on a common-language scale that non-specialists can read. Second, it derives a decision model that links the relative accuracy of human and algorithm, and the relative cost of their errors, to the value of an override. Third, it sets the experimental, field, survey and legal evidence side by side and tests how sensitive the survey association is to omitted factors.

Section 2 reviews the literature and states six hypotheses. Section 3 describes the evidence and methods. Section 4 reports the results, Section 5 discusses them, Section 6 draws implications, Section 7 sets out limitations, and Section 8 concludes.

2. Literature Review and Hypotheses

2.1. Judgment versus Algorithms

The comparison of human judgment with statistical rules is old. Meehl (1954) argued that mechanical combination of cues predicts as well as clinical judgment or better. Dawes et al. (1989) reviewed the evidence and found actuarial methods equal or superior in most studies, largely because people weight cues inconsistently and respond to irrelevant ones. Kahneman et al. (2021) call the resulting scatter noise: different people, or the same person on different days, reach different verdicts on the same case, which costs accuracy and fairness. Rules remove it. Kahneman and Klein (2009) add a condition under which intuition can be trusted, a regular environment in which the expert has had the chance to learn its regularities. Simon's (1955) bounded rationality helps explain why simple rules hold up when information and computing capacity are limited.

What has changed is the algorithm. Machine learning models use many more cues and learn nonlinear relations, and they now beat people in more tasks than they used to. They also fail in ways that a person can sometimes see: when a case lies outside the training data, when the prediction target is a poor proxy for what the organization cares about, or when the context has shifted. The value of a human reviewer is therefore the information the person holds and the model lacks, minus the noise and bias the person adds.

2.2. Aversion, Appreciation and Reliance

People respond to algorithmic advice in conflicting ways. Dietvorst et al. (2015) found that people lose confidence in an algorithm faster than in a human forecaster after seeing it err, even when the algorithm is better, a pattern they named algorithm aversion. In later experiments the same group found that people use imperfect algorithms more readily when they may adjust the output, even slightly (Dietvorst et al., 2018). Logg et al. (2019) found the opposite in other tasks: advice described as algorithmic was weighted more heavily than advice from other people, with experts the least ready to take any advice. Reliance therefore depends on the task, on how the advice is framed and on who receives it. Glikson and Woolley (2020) review the antecedents of trust in AI and reach a similar conclusion.

2.3. Automation Bias and Oversight

Bainbridge (1983) described the ironies of automation: the more reliable the system, the less practice operators get and the harder the rare cases that are left to them. Skitka et al. (1999) showed that people working with an automated aid commit omission and commission errors because they follow the aid without checking. Parasuraman and Manzey (2010) reviewed complacency and automation bias and concluded that both affect experts as well as novices and are not reliably removed by practice. In public administration, Alon-Barkat and Busuioc (2023) found selective adherence: officials followed algorithmic advice more often when it matched what they already believed. Green (2022) argues that policies requiring human oversight of government algorithms rest on faulty assumptions about what people can do, that they offer false reassurance, and that they move accountability onto individuals who have not been given the means to exercise it.

2.4. Human–AI Combinations in Experiments and Trials

Experimental work on human–AI teams has grown quickly. Vaccaro et al. (2024) pooled 106 experiments published between January 2020 and June 2023 in a preregistered meta-analysis. On average the combination did worse than the better of the human or the AI alone and better than the human alone. Losses were concentrated in decision tasks, creation tasks showed gains, and combinations did better when the human was the stronger member. Explanations and confidence displays, which have absorbed much of the design effort, did not significantly moderate performance.

Studies with expert participants point the same way. In an experiment with 227 radiologists, Agarwal et al. (2023) found that AI predictions did not improve average accuracy, because radiologists underweighted the predictions and treated their own information and the AI's as if they were independent, while contextual clinical information did improve accuracy. The authors conclude that, unless these belief-updating errors can be corrected, cases are best delegated either to the human or to the AI and rarely to the AI-assisted human. Goh et al. (2024) randomized physicians to GPT-4 or conventional resources and found no significant difference in diagnostic reasoning (median 76.3% against 73.7%), although the model working alone scored above 92%. A later trial on management reasoning did find a benefit: physicians assisted by GPT-4 scored 43.0% against 35.7% for those with conventional resources, close to the model alone at 43.7% (Goh et al., 2025). In a field experiment with 758 consultants, Dell'Acqua et al. (2023) found large gains on tasks inside the frontier of what GPT-4 could do and a 19 percentage point lower chance of a correct answer on a task outside it. Raisch and Krakowski (2021) add that automation and augmentation, often presented as alternatives, interact.

2.5. Discretion in the Field

Field studies of risk assessment in criminal justice give evidence on expert discretion. Kleinberg et al. (2018) compared the bail decisions of New York judges with a machine-learning prediction built from the same information. At the judges' own jailing rate the algorithm could cut crime by no less than 14.4% and up to 24.7%, or, with no increase in crime, cut jailing by no less than 18.5% and up to 41.9%, in part because judges released many defendants the algorithm ranked as very high risk. Stevenson and Doleac (2019) studied what happens when risk scores are handed to judges and found that judges often departed from them, so that the effect of the tool on outcomes depended on how discretion was used. The two studies show both the potential of algorithms and how much of it is lost in the way people use them.

2.6. Regulation

The European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689) requires high-risk systems to be designed so that natural persons can oversee them effectively while they are in use (Article 14). The article lists what the people assigned to oversight must be enabled to do: to remain aware of the tendency to rely too heavily on the output (automation bias), to interpret the output correctly, to decide not to use the system or to disregard, override or reverse its output, and to stop it (Article 14(4)(b) to (e)). Application dates for the high-risk obligations were first set for August 2026. The Digital Omnibus on AI was proposed by the Commission on 19 November 2025, agreed politically on 7 May 2026, approved by the Parliament on 16 June and adopted by the Council on 29 June 2026. It was published as Regulation (EU) 2026/1744 on 24 July 2026 and entered into force on 27 July 2026. It moves the application date to 2 December 2027 for stand-alone high-risk systems and to 2 August 2028 for systems embedded in regulated products. The legal summaries consulted for this paper do not report changes to the text of Article 14.

A parallel line of law sits in data protection. In the SCHUFA judgment the Court of Justice held that a credit score can itself be an automated decision under Article 22 of the General Data Protection Regulation when the party receiving it draws strongly on it, so that a nominal human sign-off does not by itself take a decision outside the rule (Court of Justice of the European Union, 2023). Whether provisions of this kind secure real oversight depends on how organizations implement them, which is the subject of this paper.

2.7. Organizational Performance

Returns from technology depend on complementary changes in process, skills and governance, and measured gains often trail adoption, as the productivity J-curve literature argues (Brynjolfsson et al., 2021). For AI decision tools the relevant complements include the allocation of authority between people and systems, the information available to those who oversee, and the feedback that lets an organization learn whether the combination works. These are the elements of oversight design, and they are the elements that organizations rarely measure.

2.8. Research Gap, Framework and Hypotheses

The literature has separate strands. Behavioral studies examine reliance, experiments examine combinations, field studies examine discretion, surveys describe organizations and legal analysis describes duties. Few studies put the evidence on one scale, state the conditions under which a human adds value, or relate oversight practices to organizational outcomes. Figure 1 summarizes the framework used here. An algorithm produces a recommendation of a certain accuracy and opacity. A human can accept it, adjust it, override it or add information. The quality of the joint decision depends on the information the human adds and the noise and bias the human introduces, and it feeds organizational performance. Oversight design, meaning who decides, with what information, authority and accountability, and under what workload, conditions the contribution of the human.

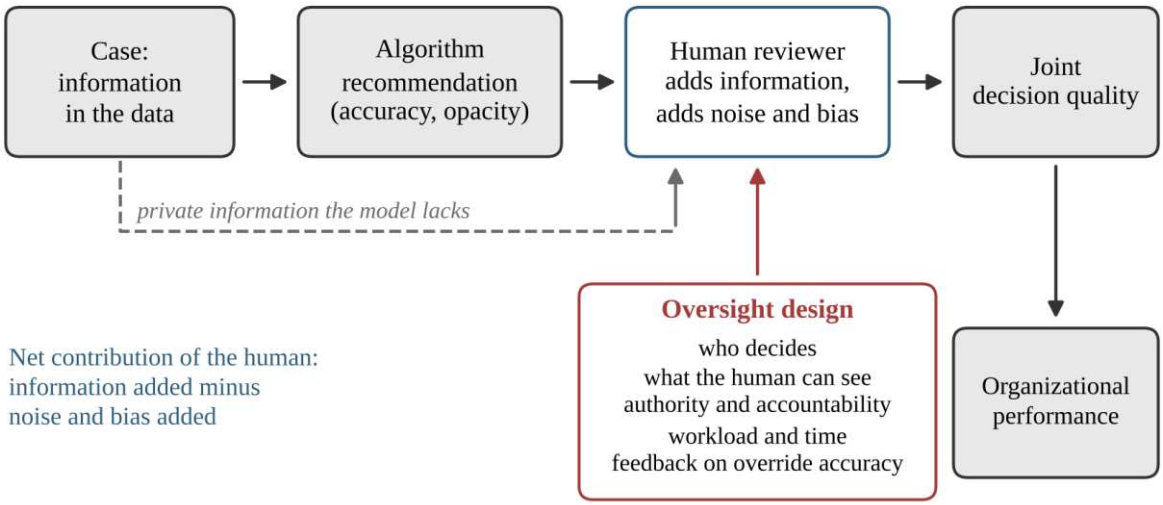

Figure 1. Conceptual framework: the human contribution depends on oversight design.

Six hypotheses follow.

H1 Human–AI combinations perform on average worse than the better of the human and the algorithm alone, and better than the human alone.

H2 Combinations perform worse in decision tasks than in creation tasks, and in decision tasks the probability that a combination outperforms its best member is below one half.

H3 Relative capability moderates the effect: combinations gain when the human alone is better than the algorithm alone and lose when the algorithm is better.

H4 A human override improves accuracy only if the human is right in more than the break-even share of overridden cases, which is one half when errors are equally costly.

H5 Field, trial and worker evidence indicates that nominal oversight cannot be assumed to be effective.

H6 Respondents at organizations with defined human-validation processes are much more likely to report high performance, while most organizations that use AI report no measurable profit impact.

H1 to H3 are tested on the meta-analytic effects, H4 by derivation, H5 on the field, trial and worker evidence, and H6 on the survey statistics.

3. Data and Methods

3.1. Research Design and Evidence Base

The study is a structured secondary synthesis of published quantitative evidence. It is not a systematic review and does not claim to be exhaustive. Sources were chosen because they report results that can be expressed on comparable scales and that bear on one of the research questions: pooled effects of human–AI combinations, field or trial evidence on expert discretion over algorithmic output, organizational survey evidence on oversight practices, and the legal text that requires oversight. The meta-analysis of Vaccaro et al. (2024) anchors the experimental evidence. It is the largest preregistered synthesis, it searched the ACM Digital Library, the Web of Science and the AIS eLibrary, screened 5,126 records, and reports pooled effects with test statistics and heterogeneity for 106 experiments and 370 effect sizes. Because it stops at June 2023, three randomized or field experiments from the generative-AI period (Dell'Acqua et al., 2023; Goh et al., 2024, 2025) were added after targeted searches in November 2026.

The organizational evidence is the 2025 McKinsey global survey, fielded from 25 June to 29 July 2025 with 1,993 participants in 105 countries (McKinsey & Company, 2025). Context comes from the U.S. Census Bureau's Business Trends and Outlook Survey, the EU-OSHA OSH Pulse 2025 worker survey (25,688 interviews in the EU27 between 31 March and 14 April 2025) and the regulatory texts. Table 1 lists the sources and how each was verified. Every figure used here was compared with the original publication or an official release in November 2026. Where only secondary summaries of a figure could be retrieved, Table 1 says so.

Table 1. Evidence base and verification status.

Source	Content used	Role in the analysis	Verification
Vaccaro et al. (2024)	Meta-analysis, 106 experiments, 370 effect sizes; pooled *g*, test statistics, moderators, robustness	H1–H3	Checked against full text
Kleinberg et al. (2018)	Bail decisions; simulated crime and jailing reductions	H5	Checked against published abstract and text
Stevenson & Doleac (2019)	Judicial use of risk scores	H5	As cited in the original draft; not re-opened
Agarwal et al. (2023)	227 radiologists; AI predictions and contextual information	H3, H5	Checked against working-paper text and abstract
Goh et al. (2024, 2025)	Randomized trials of GPT-4 with physicians	H5	Checked against journal abstract and several summaries
Dell'Acqua et al. (2023)	758 consultants; inside and outside the AI frontier	H5	Checked against working-paper abstract
McKinsey & Company (2025)	Survey of 1,993 respondents; adoption, impact, validation processes	H6	Headline figures checked; the 65% and 23% values and the 51% figure via summaries of the report
U.S. Census Bureau (2026)	Business Trends and Outlook Survey; AI use by firm size	Context	Figures matched in Census-based reports; page not re-opened
EU-OSHA (2025)	OSH Pulse 2025; digitalisation and work	H5	Checked against official infographic
Regulation (EU) 2024/1689; 2026/1744	Article 14; application dates	Section 4.7	Article 14 text checked; dates via Council and Parliament releases and legal summaries

Note. Verification was carried out in November 2026. "Summaries" are reports and news items that quote the original.

3.2. Recovering and Re-expressing Effect Sizes

For a pooled effect g with a 95% interval (l, u) and df degrees of freedom, the standard error is recovered as $SE = (u - l) / (2t^{0.975,\,df})$, using the reported degrees of freedom (92 to 104) in place of the normal quantile. The test statistics and p values quoted in the text are those reported by Vaccaro et al. (2024), who used a three-level random-effects model with cluster-robust variance estimation and the Knapp–Hartung adjustment. The recovered standard errors reproduce those statistics to within about 0.3 in t, which is consistent with rounding in the published intervals.

To give effects an interpretable scale, the probability of superiority, the chance that a randomly chosen combination outperforms a randomly chosen comparator, is computed as

$$\mathrm{CL} = \Phi(g / \sqrt{2}) \tag{1}$$

where Φ is the standard normal distribution function, which holds for normally distributed outcomes with equal variances (McGraw & Wong, 1992). Interval limits are transformed with the same function. Hedges'

g is the standardized mean difference corrected for small-sample bias (Hedges, 1981), and by convention 0.2, 0.5 and 0.8 mark small, medium and large effects (Cohen, 1988). Differences between subgroups are tested with the moderator statistics reported by the original authors. As a check, the independent-subgroups statistic $z = (g_a - g_b) / \sqrt{(SE_a^2 + SE_b^2)}$ is also computed; it is only approximate because a single study can contribute to a task-type subgroup and to a relative-performance subgroup at once.

3.3. A Decision Model for Overrides

Consider a binary decision in which an algorithm recommends an action and a human may override it. Let q be the share of all cases that the human overrides and r the share of overridden cases in which the human is right, so that the algorithm is right in the remaining $1 - r$. Overriding changes accuracy by

$$\Delta = q\,[r - (1 - r)] = q\,(2r - 1) \tag{2}$$

which is positive only if $r > 0.5$. If errors differ in cost, with c_A the cost of an algorithm error that a correct override avoids and c_H the cost of a human error introduced by an incorrect override, the expected net benefit is $q\,[r\,c_A - (1 - r)\,c_H]$ and the break-even share is

$$r^* = c_H / (c_A + c_H) \tag{3}$$

The rule links the value of an override to the human's accuracy in the cases the human chooses to override. It can be tied to the overall accuracy of the two parties. Let the human be right with probability a and the algorithm with probability b, and suppose their errors are independent given the truth. Among the cases in which they disagree, the human is right with probability

$$r = a(1 - b) / [a(1 - b) + b(1 - a)] \tag{4}$$

and disagreement occurs in a share $q = a(1 - b) + b(1 - a)$ of cases. Substituting in Equation 2 gives $\Delta = a - b$. This identity does not need the independence assumption: overriding every disagreement simply replaces the algorithm with the human, so the accuracy change is the accuracy gap. Independence is needed only for Equation 4. With unequal costs, an override pays when $a(1 - b)\,c_A > (1 - a)\,b\,c_H$, that is, when the human's odds of being right exceed the algorithm's odds by at least the cost ratio $\kappa = c_H / c_A$. Solving for the human accuracy needed,

$$a^* = \kappa b / (\kappa b + 1 - b) \tag{5}$$

The model is a benchmark and not an estimate. It treats decisions as binary, assumes that a disagreement implies that exactly one party is right, and ignores learning, trust dynamics and strategic behavior.

3.4. Oversight and Performance in the Survey

McKinsey classifies respondents as AI high performers when they attribute 5% or more of enterprise EBIT to AI and report significant value from it, about 6% of respondents. Among high performers, 65% say their organization has defined processes for when model outputs need human validation, against 23% of other respondents. Because the report gives percentages and not group sizes, the number of high performers is approximated as $n_1 = 120$ (6% of 1,993), with 100 to 140 as a sensitivity range, and $n_0 = 1{,}993 - n_1$. The risk ratio of having a validation process, high performers relative to others, is $RR = p_1 / p_0$ with log-scale variance $(1 - p_1) / (p_1 n_1) + (1 - p_0) / (p_0 n_0)$ (Katz et al., 1978), and the odds ratio is $OR = [p_1 / (1 - p_1)] / [p_0 / (1 - p_0)]$ with variance $1 / [n_1 p_1 (1 - p_1)] + 1 / [n_0 p_0 (1 - p_0)]$.

Bayes' rule converts the survey percentages into the probability of being a high performer given a validation process. With $P(H)$ the share of high performers and V a validation process, $P(V) = P(H)\,P(V|H) + [1 - P(H)]\,P(V|\text{not } H)$ and $P(H|V) = P(H)\,P(V|H) / P(V)$; $P(H|\text{not } V)$ follows in the same way. A counterfactual share of high performers at prevalence v of validation is $v\,P(H|V) + (1 - v)\,P(H|\text{not } V)$. It assumes a causal and unconfounded association and serves only as an upper-bound illustration. Intervals come from 200,000

Monte Carlo draws, with the high-performer group size drawn uniformly from 100 to 140 and binomial sampling of the two percentages (Rubinstein & Kroese, 2017).

To judge how easily an omitted factor could produce the association, the E-value of VanderWeele and Ding (2017) is computed, $E = RR + \sqrt{[RR(RR - 1)]}$. It is the minimum risk-ratio association that an unmeasured confounder would need with both validation and high performance to explain away the observed ratio. It says nothing about measurement error or reverse causation.

3.5. Verification and Reproducibility

All derived values were computed in Python with NumPy and SciPy from the published inputs in Tables 2 and 6 and the equations above, with a fixed random seed for the simulations. The recomputed risk ratio, odds ratio, Bayes shares and Monte Carlo intervals match the values reported in this paper to the stated precision. Study-level data from the meta-analysis are public through the authors' Open Science Framework repository, but they were not re-analyzed here, so the pooled estimates and moderator tests are taken as published.

4. Results

4.1. Human–AI Combinations against the Best Member

Table 2 and Figure 2 give the pooled effects as reported by Vaccaro et al. (2024). Against the better of the human or the AI alone, the pooled effect was negative: $g = -0.23$, 95% CI −0.39 to −0.07, $t(92) = -2.89$, $p = .005$. On the common-language scale, a randomly chosen combination outperforms its best member in 43.5% of comparisons (39.1% to 48.0%). Against the human alone the combination does clearly better, $g = 0.64$ (0.53 to 0.74), which corresponds to a 67.5% probability of superiority. H1 is supported. Using AI alongside people beats leaving people unaided, but on average it is worse than letting the stronger party work alone.

Table 2. Pooled effects of human–AI combinations (Vaccaro et al., 2024) and common-language equivalents.

Comparison	*g* (95% CI)	*SE*	*t* (*df*)	*p*	Probability combination outperforms, % (95% CI)
Combination vs best of human or AI alone (synergy)	−0.23 (−0.39, −0.07)	0.081	−2.89 (92)	.005	43.5 (39.1–48.0)
Combination vs human alone (augmentation)	0.64 (0.53, 0.74)	0.053	11.87 (98)	<.001	67.5 (64.6–70.0)
Synergy, decision tasks	−0.27 (−0.44, −0.10)	0.086	−3.20 (104)	.002	42.4 (37.8–47.2)
Synergy, creation tasks	0.19 (−0.09, 0.48)	0.144	1.35 (104)	.180	55.3 (47.5–63.3)
Synergy, human alone better than AI alone	0.46 (0.28, 0.66)	0.096	5.06 (104)	<.001	62.8 (57.8–68.0)
Synergy, AI alone better than human alone	−0.54 (−0.71, −0.37)	0.086	−6.20 (104)	<.001	35.1 (30.8–39.7)
Augmentation, AI alone better than human alone	0.74 (0.63, 0.85)	0.055	13.50 (104)	<.001	70.0 (67.2–72.6)

Note. *g*, *CI*, *t*, *df* and *p* are as reported by Vaccaro et al. (2024). *SE* is recovered from the interval with the *t* quantile at the reported *df* (Section 3.2). Probabilities of superiority are $\Phi(g/\sqrt{2})$ (Equation 1). Synergy rows compare the combination with the better of the two members.

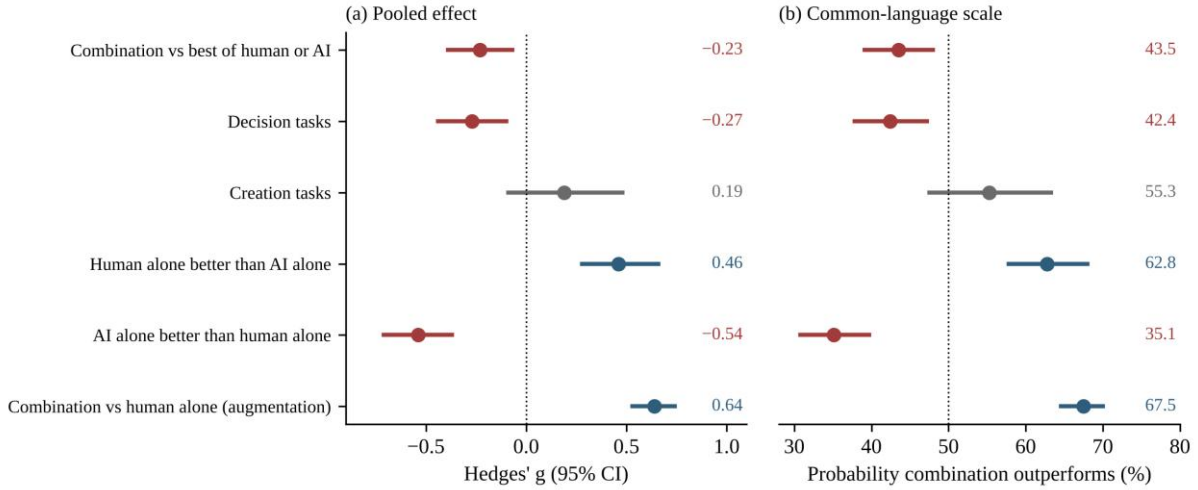

Figure 2. Pooled effects of human–AI combinations: (a) Hedges' *g* with 95% CI; (b) probability that the combination outperforms the comparator.

Source: Vaccaro et al. (2024); probabilities computed with Equation 1. Blue and red mark significant positive and negative effects; gray marks an interval that includes no effect.

Two features of the evidence keep this from being read as a law of nature. The first is heterogeneity. The original authors report $I^2 = 97.7\%$ for the synergy effect (Higgins & Thompson, 2002), so nearly all of the variation among effect sizes reflects real differences between settings and not sampling error. The pooled mean is a poor guide to what any one organization will find. The second is robustness. The result does not depend on a few studies or on the clustering choice. Clustering at the paper level gave $g = -0.22$ (−0.41 to −0.04), dropping 11 outliers gave $g = -0.25$ (−0.39 to −0.11), and leave-one-out estimates ran from −0.28 to −0.19, all with $p < .05$ (Table 3, panel B). Egger's regression showed no sign of publication bias for the synergy outcome ($p = .438$) but did for the comparison with humans alone ($p = .002$), so the more flattering comparison is also the one more likely to be overstated.

In standardized terms the loss is small by Cohen's (1988) benchmarks. In operational terms a shortfall of 6.5 percentage points in the probability of superiority, 43.5 against 50, is not trivial when it applies to every case in a high-volume process such as credit approval or claims handling.

4.2. Task Type

The loss is concentrated in decision tasks, defined by Vaccaro et al. as tasks in which participants choose among a finite set of options: $g = -0.27$ (−0.44 to −0.10), $t(104) = -3.20$, $p = .002$, a probability of superiority of 42.4% (37.8% to 47.2%). In creation tasks, where participants produce open-ended content, the estimate is positive but imprecise: $g = 0.19$ (−0.09 to 0.48), $t(104) = 1.35$, $p = .18$, or 55.3% (47.5% to 63.3%). That estimate rests on only 34 effect sizes, so the lack of significance should not be read as a lack of gain. The difference between the two task types is 0.46 and is statistically significant, $F(1, 104) = 7.84$, $p = .006$ (Table 3, panel A). H2 is supported: in decision tasks the probability that a combination beats its best member is below one half, and so is the upper confidence limit.

About 85% of the effect sizes in the meta-analysis come from decision tasks and about 10% from creation tasks (Vaccaro et al., 2024). The literature therefore speaks mainly about the setting in which oversight laws bite hardest and says least about content creation with generative models. The original authors suggest that creation tasks contain routine filling-in that AI does as well as people alongside parts that need human insight, whereas in decision tasks both parties usually make a complete decision and the human has the last word. Over 95% of the combinations they analyzed had humans making the final decision, which is the standard design of human-in-the-loop systems. In the three experiments that assigned subtasks in advance, the pooled estimate was positive ($g = 0.22$) but very imprecise (−0.42 to 0.87), too uncertain to carry weight on its own.

Table 3. Moderator tests and robustness checks.

Panel A. Moderator contrasts

Moderator	Contrast	Difference in *g*	Reported test	*p*	Approx. 95% CI
Task type	Creation minus decision	0.46	$F(1, 104) = 7.84$	.006	0.13 to 0.79
Relative performance	Human better minus AI better	1.00	$F(1, 104) = 81.79$	<.001	0.78 to 1.22

Panel B. Robustness and heterogeneity (synergy outcome)

Check	*g* (95% CI)	Test
Primary model, clustered by experiment	−0.23 (−0.39, −0.07)	$t(92) = -2.89$, $p = .005$
Clustered by paper	−0.22 (−0.41, −0.04)	$p = .017$
Excluding 11 outliers	−0.25 (−0.39, −0.11)	$p = .001$
Leave-one-out (effect size, experiment, paper)	−0.28 to −0.19	all $p < .05$
Egger regression, synergy	β = −0.67 (−2.39, 1.04)	$p = .438$
Egger regression, augmentation vs human alone	β = 1.96 (0.76, 3.16)	$p = .002$
Heterogeneity, I^2	97.7% (synergy); 93.8% (augmentation)	

Note. Panel A: the difference is computed from the pooled subgroup estimates; the test statistic is the moderator test reported by Vaccaro et al. (2024); the interval uses the standard error implied by the test, difference/$\sqrt{F}$, with the *t* quantile at 104 *df*. An independent-subgroups approximation gives 0.13 to 0.79 and 0.75 to 1.25. Panel B reports checks described by Vaccaro et al. (2024).

4.3. Relative Capability

The sign of the effect follows the ranking of the two parties. When the human alone outperformed the AI alone, the combination gained: $g = 0.46$ (0.28 to 0.66), $t(104) = 5.06$, a probability of superiority of 62.8% (57.8% to 68.0%). When the AI alone outperformed the human, the combination lost: $g = -0.54$ (−0.71 to −0.37), $t(104) = -6.20$, or 35.1% (30.8% to 39.7%). The gap between the two estimates is 1.00 standard deviations. The moderator test reported by the original authors is $F(1, 104) = 81.79$, $p < .001$. It implies a standard error of the difference near 0.11 and a 95% interval of about 0.78 to 1.22. Treating the two subgroups as independent gives 0.75 to 1.25 and $z = 7.7$, which is reported only as a check. H3 is supported. Relative capability is the largest moderator in the study, larger than task type.

Two cautions apply. The moderator is defined from outcomes inside the same experiments, since each effect is classified by which baseline did better, so it describes where combinations help and does not by itself explain why. And when the AI was the stronger party the combination still beat the human alone ($g = 0.74$, 0.63 to 0.85), so such systems help the person even as they fall short of the machine.

A pair of experiments reviewed in the meta-analysis shows the pattern in miniature. In the same interface and with an AI that was 73% accurate in both tasks, participants detecting fake hotel reviews scored 55% alone and 69% with the AI, below the AI alone, while participants classifying bird images scored 81% alone and 90% with the AI, above both (Cabrera et al., 2023, as described in Vaccaro et al., 2024). The radiology experiment gives a mechanism. Radiologists underweighted the AI's predictions and treated them as independent of their own information, and delegation to the better party did better than assistance (Agarwal et al., 2023). People who do not know whether they or the model is more accurate in a given case err in both directions (Dietvorst et al., 2015; Logg et al., 2019; Parasuraman & Manzey, 2010).

4.4. When Does an Override Pay?

Under Equation 2 an override changes accuracy by $q(2r - 1)$ percentage points (Figure 3; Table 4, panel A). If humans override 10% of cases and are right in half of them, accuracy is unchanged. At $r = 0.4$ it falls by

2 points and at $r = 0.6$ it rises by 2. At a 20% override rate the swings are 4 points at $r = 0.4$ or 0.6 and 8 points at $r = 0.3$ or 0.7. A high override rate is therefore not a virtue in itself. When r is below the break-even level, a larger q only means larger losses. With unequal costs the break-even moves: if a human error costs twice as much as an algorithm error, the human must be right in two thirds of overridden cases, and if an algorithm error costs twice as much, one third is enough. H4 is supported by derivation.

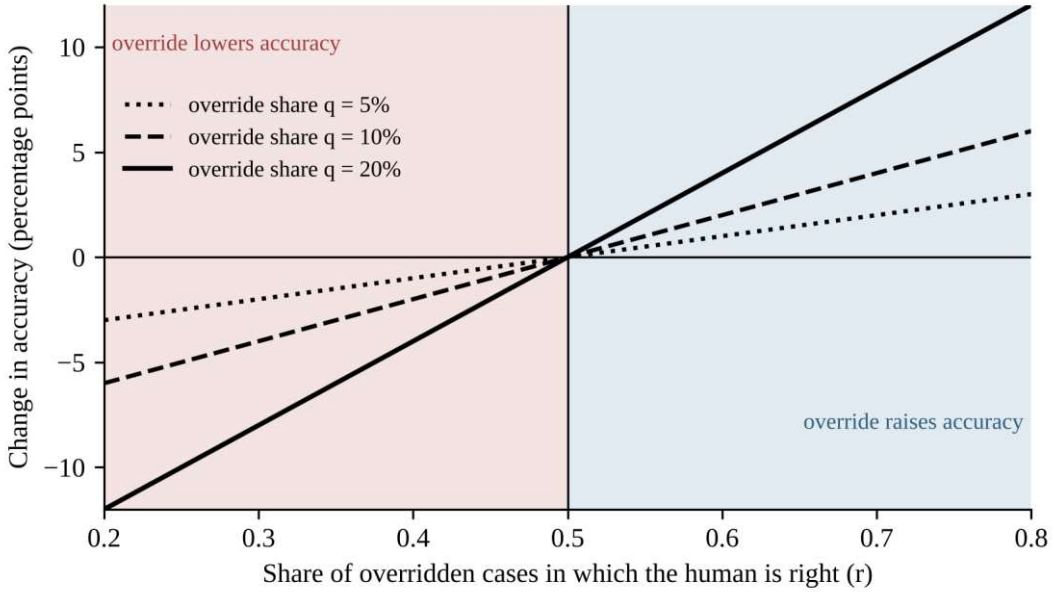

Figure 3. Change in accuracy from human overrides, $\Delta = q(2r - 1)$.
Break-even is $r = 0.5$ when errors are equally costly.

The identity $\Delta = a - b$ (Section 3.3) ties the rule to the meta-analytic finding. If an organization lets people override every disagreement, accuracy moves by the gap between human and algorithm accuracy, whatever the pattern of errors. If the algorithm is 80% accurate and the human 70%, full discretion costs 10 points; if the human is 85%, it gains 5 (Table 4, panel C). Under independent errors the humans in the first case are right in only 36.8% of the cases they override. The condition for a gain is stricter when human errors are costlier. At 80% algorithm accuracy and equal costs the human must exceed 80%; if a human error costs twice as much, 88.9%; if it costs four times as much, 94.1% (Equation 5; Table 4, panel B; Figure 4).

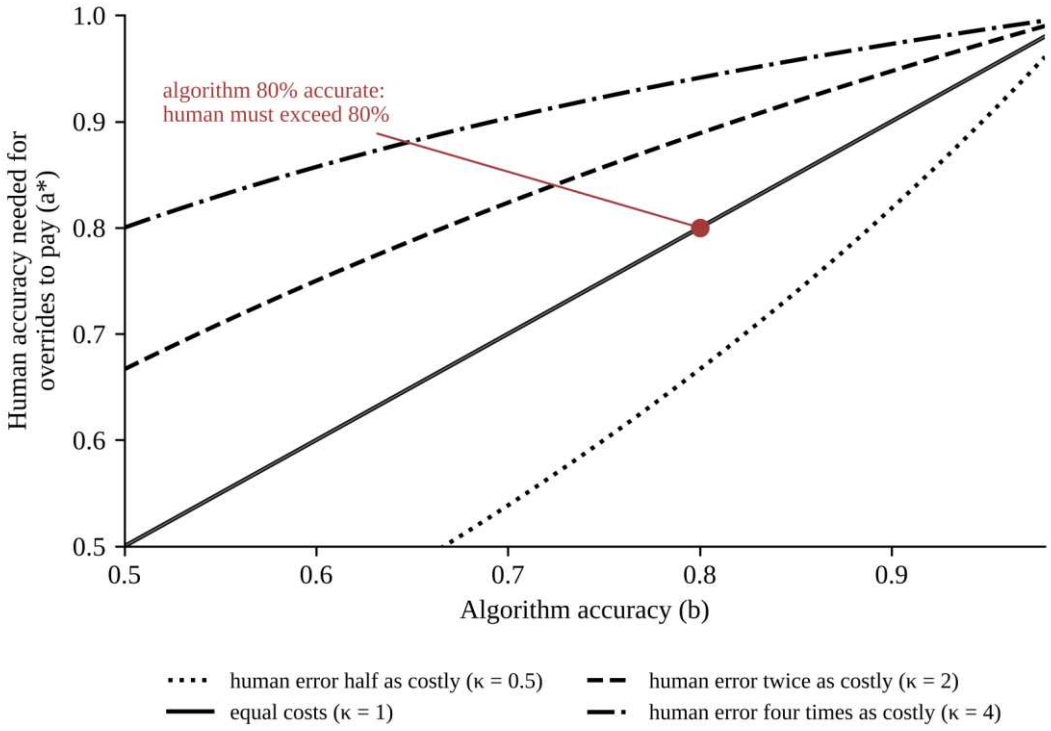

Figure 4. Human accuracy needed for overrides to pay, by algorithm accuracy and cost ratio $\kappa = c_H/c_A$. Equation 5.

Table 4. Decision model for human overrides.

Panel A. Change in accuracy (percentage points), $\Delta = q(2r - 1)$

Override share (q)	$r = 0.3$	$r = 0.4$	$r = 0.5$	$r = 0.6$	$r = 0.7$
5%	−2.0	−1.0	0.0	1.0	2.0
10%	−4.0	−2.0	0.0	2.0	4.0
20%	−8.0	−4.0	0.0	4.0	8.0

Panel B. Human accuracy needed for overrides to pay, $a^* = \kappa b / (\kappa b + 1 - b)$

Algorithm accuracy (b)	$\kappa = 0.5$	$\kappa = 1$	$\kappa = 2$	$\kappa = 4$
0.70	0.538	0.700	0.824	0.903
0.80	0.667	0.800	0.889	0.941
0.90	0.818	0.900	0.947	0.973
0.95	0.905	0.950	0.974	0.987

Panel C. Unrestricted override with independent errors

Human accuracy (a)	Algorithm accuracy (b)	Disagreement share (q)	Human right when overriding (r)	Δ (points)
0.70	0.80	0.380	0.368	−10.0
0.80	0.80	0.320	0.500	0.0
0.85	0.80	0.290	0.586	5.0
0.60	0.90	0.420	0.143	−30.0
0.90	0.90	0.180	0.500	0.0

Note. Break-even share $r^* = c_H/(c_A + c_H)$: 0.50 with equal costs, 0.33 if an algorithm error costs twice a human error, 0.67 if a human error costs twice an algorithm error, 0.75 if it costs three times. $\kappa = c_H/c_A$. Panel C uses Equation 4; $\Delta = a - b$ holds in general for full override.

Two consequences follow. When the algorithm is the stronger party, overrides are likely to have r below one half and the combination loses to the algorithm alone, as the effect of −0.54 shows. When the human has information the model lacks, overrides concentrated on those cases have r above one half and the combination gains. What matters, then, is not whether a human is in the loop but whether the human's interventions fall where the human has an informational advantage. Selective override can do better than the $a - b$ that full discretion delivers, whereas full discretion helps only when the human is the stronger party.

4.5. Field, Trial and Worker Evidence

Table 5 sets the field and trial studies side by side. They show three patterns. First, when the algorithm is the stronger predictor and the human sees mostly the same information, overrides add little. In the New York bail data the algorithm could have cut crime by 14.4% to 24.7% at the judges' jailing rate, in part because judges released defendants the algorithm ranked as very high risk, and a model trained to predict the judges' own decisions did better than the judges, which the authors read as judges responding to noise as if it were signal (Kleinberg et al., 2018). In the physician trial with GPT-4 the model alone scored above 92% and physicians with and without it scored about 74% to 76% (Goh et al., 2024). Second, where humans hold information the model lacks, using it helps: in the radiology experiment contextual information improved accuracy while AI predictions did not (Agarwal et al., 2023). Third, the value of assistance depends on whether the task lies inside the model's competence and whether users can tell. Consultants gained on tasks inside the frontier and were 19 percentage points less likely to be correct on the task outside it (Dell'Acqua et al., 2023).

Table 5. Field and trial evidence on human oversight of algorithmic output.

Study	Setting and design	Main finding	Reading for oversight design
Kleinberg et al. (2018)	New York bail decisions; machine-learning prediction vs judges	At the judges' jailing rate, crime could fall by 14.4% to 24.7%; or jailing by 18.5% to 41.9% at constant crime	Overrides added little where the algorithm predicted better
Stevenson & Doleac (2019)	Risk scores supplied to judges	Judges often departed from the scores; effects depended on use of discretion	Discretion shapes the value of the tool
Agarwal et al. (2023)	227 radiologists; chest X-rays; AI predictions and contextual information varied	AI predictions did not raise average accuracy; contextual information did; AI underweighted	Delegate or give information; assistance alone is weak
Goh et al. (2024)	Randomized trial, physicians, diagnostic vignettes	Median 76.3% with GPT-4 vs 73.7% conventional (not significant); GPT-4 alone above 92%	Assistance without synergy
Goh et al. (2025)	Randomized trial, 92 physicians, management reasoning	43.0% with GPT-4 vs 35.7% conventional; GPT-4 alone 43.7%	Gain vs unaided, no gain vs AI alone
Dell'Acqua et al. (2023)	758 consultants; tasks inside and outside the AI frontier	Inside: 12.2% more tasks, 25.1% faster, higher quality. Outside: 19 percentage points less likely correct	Value depends on task fit and on users recognizing it
Dietvorst et al. (2018)	Experiments on use of imperfect algorithms	Allowing small adjustments raised use of the algorithm	Limited adjustment can raise adoption

Note. Figures are as reported in the sources. The Goh et al. (2025) values are as reported in summaries of the article.

Two design options follow. One is delegation: assign each case, or class of cases, to the human or to the algorithm according to which has performed better on it, as the radiology study suggests. The other is structured escalation: let the algorithm decide by default and ask a human to review only cases that meet defined conditions, such as low model confidence or inputs outside the training data. Both need evidence on relative performance, which organizations can collect by logging recommendations, overrides and

outcomes. Dietvorst et al. (2018) offer a third, softer route: giving people limited authority to adjust the output raised their willingness to use the algorithm.

The oversight problem has a counterpart on the worker side. In the EU-OSHA OSH Pulse 2025, 48% of EU workers said that digital technologies determined the speed or pace of their work, 28% that they increased workload, 19% that they reduced opportunities to use their knowledge and skills, and 16% that they reduced their ability to decide how to do their work (EU-OSHA, 2025). Oversight takes time, information and authority. Workers whose pace is set by a system and whose workload is raised by it have less of all three. H5 is supported as far as these data go: nominal oversight cannot be assumed to be effective when the conditions for exercising it are eroding. The survey does not say how many of the workers concerned oversee algorithmic decisions, so the inference is about conditions and not about outcomes.

4.6. Oversight Processes and Organizational Performance

Table 6 and Figure 5 give the organizational evidence. In the McKinsey survey 88% of respondents said their organization used AI in at least one function, but 39% attributed any enterprise-level EBIT impact to it, which is 44% of those reporting use. About 7% said AI was fully scaled and about 6% met the high-performer definition (EBIT impact of 5% or more and significant value). About 51% reported at least one negative consequence of AI use (McKinsey & Company, 2025). Among U.S. employer firms, 19.8% reported using AI in a business function in the period ending 3 May 2026, and 37% of firms with 250 or more employees did (U.S. Census Bureau, 2026). The Census question was widened in November 2025 from AI use in producing goods or services to use in any business function, so the series is not comparable across that date, and later biweekly releases show further growth. Adoption is much wider than measurable benefit, and the benefit is concentrated.

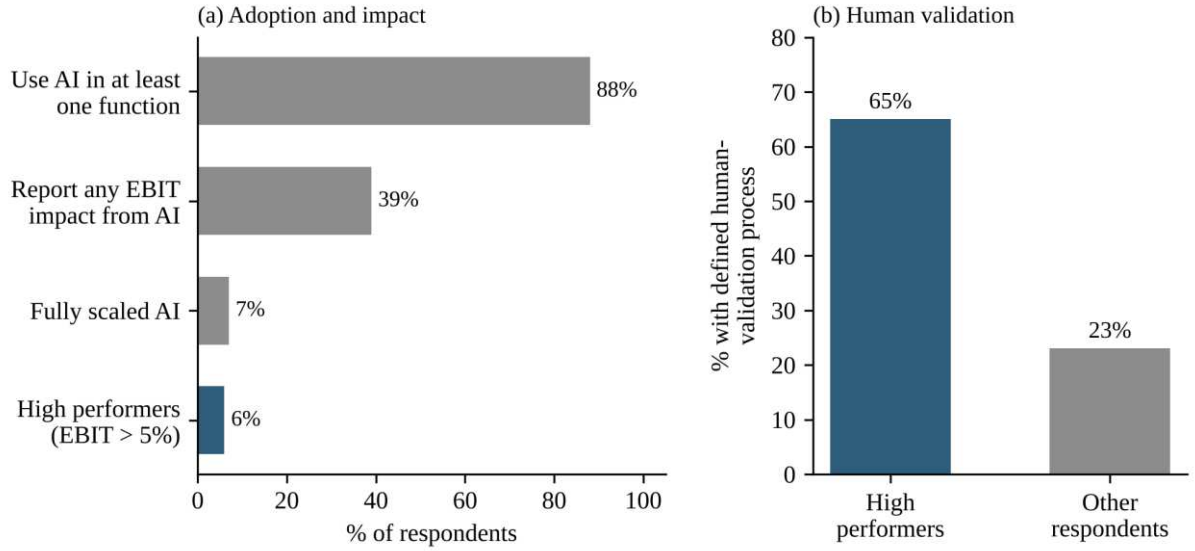

Figure 5. McKinsey (2025) survey: (a) adoption and impact; (b) defined human-validation processes.
Percentages of respondents. Values for the "fully scaled" and validation panels are as reported in summaries of the survey.

Among high performers, 65% reported defined processes for when model outputs need human validation, against 23% of the others. The risk ratio is 2.83 (95% CI 2.42 to 3.30; 2.39 to 3.34 with a high-performer group of 100 and 2.44 to 3.27 with 140) and the odds ratio is 6.22 (4.21 to 9.19). H6 is supported. Applying Bayes' rule, the implied prevalence of a validation process is 25.5%. The share of high performers is 15.3% among respondents whose organizations have one (Monte Carlo interval 12.2% to 18.6%) and 2.8% among those without (2.0% to 3.7%), a ratio of 5.4 (3.8 to 8.0).

These numbers need careful reading. The association is unadjusted and cross-sectional. High performers may have more mature governance in general, and a validation process may mark competence and not cause

it. The E-value for the ratio of 5.4 is 10.3, and 7.1 for its lower interval limit. An unmeasured factor would therefore need a risk-ratio association of about 7 or more with both validation and high performance to move the interval to the null. That is a strong requirement, though general management quality is a plausible candidate for something strongly related to both. The E-value does not address measurement error, common-method bias (both items come from the same respondent, who must also judge what share of EBIT is due to AI) or reverse causation, since organizations that already see value from AI may be the ones that formalize validation.

If the association were wholly causal and prevalence rose to 65%, the share of high performers would rise from 6.0% to 10.9% (8.9% to 13.1%), and at universal adoption to 15.3% (Figure 6). That is an upper bound. Its purpose is to give a sense of scale and not to forecast. The pattern fits the logic of Section 4.4. A process that specifies when outputs need human validation directs human attention to defined cases and so concentrates overrides where they are likely to be informed, instead of spreading oversight thinly over everything. It also makes accountability explicit. That three quarters of respondents report no such process suggests that oversight is more often assumed than designed.

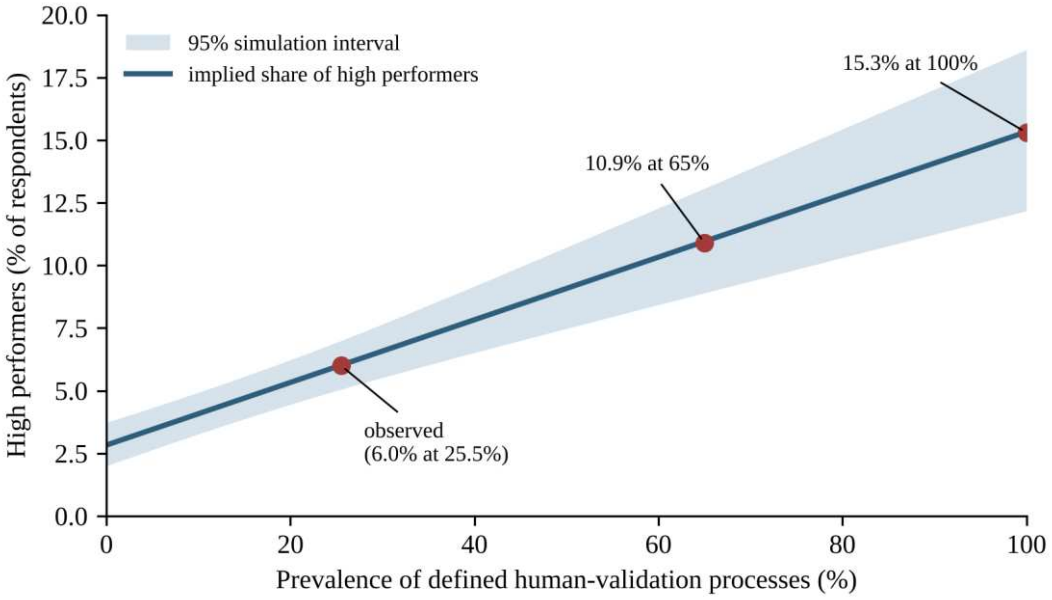

Figure 6. Counterfactual share of high performers as the prevalence of human-validation processes rises.
Upper-bound illustration that assumes the association is causal and unconfounded; 95% simulation interval from 200,000 draws.

4.7. The Regulatory Setting

The European Union has made human oversight a legal requirement for high-risk systems. Article 14(4) goes further than requiring a person in the loop: it names automation bias, correct interpretation of the output, the ability to disregard, override or reverse it, and the ability to stop the system as things the overseer must be enabled to do. The Digital Omnibus changes when these duties apply (Table 7) and not, according to the summaries consulted, what they say. The evidence in this paper suggests that compliance by presence, with a person formally in the loop, would not deliver the outcomes the requirement intends. The conditions for effective oversight include information the human has and the model lacks, authority to act on it, a workload that allows attention, and evidence on override accuracy. The SCHUFA judgment points the same way from the data protection side, since a decision does not stop being automated because a person signs it off.

Table 7. Adoption, worker conditions and regulatory timeline.

Panel A. Adoption and worker conditions

Measure	Value	Source
U.S. employer firms using AI in a business function (period ending 3 May 2026)	19.8%	U.S. Census Bureau (2026)
Firms with 250 or more employees	37%	U.S. Census Bureau (2026)
EU workers: digital technologies determine pace of work	48%	EU-OSHA (2025)
…increase their workload	28%	EU-OSHA (2025)
…reduce opportunities to use knowledge and skills	19%	EU-OSHA (2025)
…reduce ability to decide how to do the work	16%	EU-OSHA (2025)

Panel B. AI Act timeline

Date	Event
13 June 2024	Regulation (EU) 2024/1689 adopted; Article 14 requires effective human oversight of high-risk systems
19 November 2025	Commission proposes the Digital Omnibus on AI
7 May 2026	Political agreement between Council and Parliament
16 June / 29 June 2026	Approval by the Parliament / adoption by the Council
24 July / 27 July 2026	Publication as Regulation (EU) 2026/1744 / entry into force
2 December 2027	New application date, stand-alone high-risk systems
2 August 2028	New application date, high-risk systems embedded in regulated products

Note. Census figures are for the question wording introduced in November 2025 and are not comparable with earlier releases. Dates for the Digital Omnibus are from official releases and legal summaries and should be confirmed against the Official Journal text.

5. Discussion

The study began from a premise common in organizations and in law, that a human in the loop makes algorithmic decisions better. The published evidence supports a more qualified view. On average a human–AI combination beats the human and loses to the better of the two, so the human adds something only when the human is the stronger party or holds information the model lacks. In decision tasks, the kind that organizations most often automate, the chance that a combination beats its best member is about 42%. The reasons are those the behavioral literature has described for decades. People over- and under-trust algorithms for reasons unrelated to accuracy, they grow complacent, and they often do not know whether they or the model is better on the case in front of them.

The decision model puts the matter in terms an organization can measure. An override helps only if the human is right in more than the break-even share of the cases overridden, and with independent errors that condition collapses into a comparison of the human's and the algorithm's odds of being right, scaled by the relative cost of their errors. Every override is a bet on the human's accuracy in that case. Organizations that count how often people override, and how often they are right when they do, can tell whether oversight is adding value and where. Those that count only whether a human was in the loop cannot. The finding that explanations and confidence displays did not significantly move performance in the meta-analysis fits this reading: telling a reviewer more about the model helps little if the reviewer cannot judge when to trust it.

The newer trials add a complication and a caution. They add a complication because the generative-AI period is not simply a repeat of the earlier one. Physicians assisted by GPT-4 did better than physicians without it on management reasoning, though no better than the model alone (Goh et al., 2025), and consultants gained substantially inside the model's frontier (Dell'Acqua et al., 2023). The caution is that the

same studies show assistance without synergy in diagnosis (Goh et al., 2024) and a clear loss outside the frontier, where the AI-assisted answers were also rated more persuasive. Because the edge of that frontier is invisible to the user, the question of when to defer is hard to answer from inside the task, which is the situation of the radiologists who underweighted the model.

The survey evidence suggests that a useful practice is also a rare one. About a quarter of respondents report defined processes for when outputs need human validation, and the odds of high performance are six times higher among those that do. The association is not causal, and the sensitivity analysis shows only that a single omitted factor would have to be strong to explain it. It still agrees with the logic of directing human attention to defined cases. It also sits against a wide gap between adoption and impact: 88% report use, 39% any profit effect, 6% high performance. Many organizations use AI without the governance that would let them tell whether it helps.

Accountability needs a separate word. Where the law or an organization requires an accountable person to take the decision, the question is not whether the person is more accurate than the model but whether the person can exercise real judgment. Someone who endorses most recommendations within minutes is accountable in name. The evidence here suggests that organizations should say which objective they are serving, accuracy or accountability, and design oversight for it, because a process built for one may not deliver the other. Where accuracy is the aim, the case for delegation or structured escalation is strong. Where accountability is the aim, the organization has to supply the time, information and authority that make the person's judgment real, and accept that this may cost accuracy.

For regulation, the implication is that the oversight requirement in the AI Act will be effective only if it is implemented as a design problem. Having a person nominally responsible does not secure oversight if that person lacks time, information or authority, as the worker evidence suggests (EU-OSHA, 2025), and it may shift blame onto an individual who could not have prevented the error (Green, 2022). The Act's own wording, which asks that overseers be enabled to remain aware of automation bias and to override, shows that the drafters saw part of the problem. What the text leaves to implementers is the part this paper's evidence says matters most: where in the process the human adds information.

6. Implications

6.1. For Organizations

Six practices follow from the evidence. First, decide task by task who decides, using measured relative performance and not a general belief that people or models are better. Where the model is clearly stronger and the human sees the same information, delegate to the model and keep people for exceptions. Second, give the human information the model lacks and the authority to act on it, since a reviewer who sees only the model's output has little to add to it. Third, define when validation is required, tying it to low model confidence, inputs outside the training data or a high cost of error, so that human attention goes where it can change the answer. Fourth, measure override rates and override accuracy against the break-even share, which requires logging recommendations, overrides and outcomes. Fifth, allow limited, structured adjustment of the algorithm's output, which can raise willingness to use a good model (Dietvorst et al., 2018). Sixth, protect the conditions for oversight: time, training and the authority to disagree.

6.2. For Regulators and Standard-Setters

Oversight requirements should be written and audited in terms of capability and evidence, not presence: what the human can see, whether the human can act on it, and whether the interventions improve outcomes. Reporting override rates and override accuracy for high-risk deployments would turn a formal duty into something that can be checked. Standards could also ask deployers to document, for each use, why the human is expected to know something the system does not.

6.3. For Research

Three priorities stand out. Field experiments that vary oversight design, for example delegation against assistance against escalation, would give the causal evidence the present synthesis cannot. Studies in organizations that measure override accuracy directly would test the break-even rule where it matters. And the experimental literature needs to catch up with current systems: the meta-analysis stops in mid-2023, its year-of-publication moderator was significant, and the study-level data are public. A re-analysis that adds 2023–2026 experiments and fits the three-level model with relative capability, task type and process design as joint moderators is the obvious next step.

7. Limitations

The limitations are considerable. First, the experiments are laboratory and online tasks in which participants have limited stakes, training and accountability, and the pooled effect averages over very different tasks, so it does not predict what a given organization will find. Heterogeneity is extreme (I^2 = 97.7%). Most experiments predate current generative models. The probability of superiority is a transformation that assumes normal outcomes with equal variances. Second, subgroup contrasts overlap, the moderators are not mutually adjusted, and the relative-capability moderator is defined from the outcomes themselves. This paper uses the published pooled results and moderator tests and did not re-analyze study-level data. Third, the decision model is a benchmark. It treats decisions as binary, assumes that disagreement implies one correct answer, and leaves out learning, trust dynamics and strategic behavior; the expression for r also relies on independent errors.

Fourth, the survey is a self-report. Respondents are individuals at all levels of organizations, and the same person reports both the validation process and the share of EBIT due to AI. The size of the high-performer group is approximated from a rounded percentage, some figures were retrieved from summaries of the report and not from its exhibits, and the association between validation processes and performance is unadjusted and cannot be read as causal. The E-value speaks to omitted confounders only. Fifth, the field and trial studies concern particular institutions, tasks and periods, and the trials added from the generative-AI period are few. Sixth, accuracy is not the only objective in many decisions. Fairness, explanation and legitimacy matter, and a human decision-maker may be required for reasons that an accuracy comparison cannot capture. Finally, the regulatory account rests on official releases and legal summaries of the Digital Omnibus and on the published text of Article 14, and the final Official Journal text should be consulted for any compliance purpose.

8. Conclusion

This paper asked when human judgment improves algorithmic decisions and how oversight relates to organizational performance. Across 106 experiments the pooled combination performs worse than the better of its members (g = −0.23; probability of outperforming 43.5%), especially in decision tasks (g = −0.27; 42.4%) and when the algorithm is the stronger party (g = −0.54), and better than the human alone (g = 0.64). A simple model shows that an override improves accuracy only if the human is right in more than the break-even share of the cases overridden, and that unrestricted discretion changes accuracy by exactly the gap between human and algorithm accuracy. In a global survey, 88% of respondents reported AI use and 39% any profit impact, and high performers were 2.8 times as likely to report defined processes for when outputs need human validation (odds ratio 6.2).

The limits are those of pooled published results, approximated group sizes and cross-sectional self-reports. Future work should identify the causal effect of oversight design in organizations, measure override accuracy directly, extend the meta-analytic evidence to current systems, and follow how the human oversight requirements are implemented. Until then the safest conclusion is that human judgment is valuable

where it adds information and costly where it only adds noise, and that the task of oversight design is to tell the two apart.

Author Contributions

[To be completed by the author. For a single-author paper: Conceptualization, methodology, formal analysis, writing—original draft preparation, and writing—review and editing were carried out by the author, who has read and agreed to the published version of the manuscript.]

Declaration of Generative AI and AI-Assisted Technologies in Manuscript Preparation

During revision of this manuscript an AI assistant (Claude, Anthropic) was used to check published figures against their sources, to recompute derived statistics and to edit the language of the text. The author reviewed all output and takes full responsibility for the content. [Author to confirm or amend this statement to match the journal's policy and actual use.]

Data Availability

All data are from public sources listed in the reference list. Derived calculations can be reproduced from Tables 2 and 6 and the equations in Section 3. Study-level data for the meta-analysis of Vaccaro et al. (2024) are available through the authors' Open Science Framework repository. No new primary data were collected.

Funding

This study received no external funding.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Agarwal, N., Moehring, A., Rajpurkar, P., & Salz, T. (2023). *Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology* (NBER Working Paper No. 31422, revised November 2025). National Bureau of Economic Research. https://doi.org/10.3386/w31422

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice. *Journal of Public Administration Research and Theory, 33*, 153-169. https://doi.org/10.1093/jopart/muac007

Bainbridge, L. (1983). Ironies of Automation. *Automatica, 19*, 775-779. https://doi.org/10.1016/0005-1098(83)90046-8

Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The Productivity J-Curve: How Intangibles Complement General Purpose Technologies. *American Economic Journal: Macroeconomics, 13*, 333-372. https://doi.org/10.1257/mac.20180386

Cabrera, Á. A., Perer, A., & Hong, J. I. (2023). Improving Human–AI Collaboration with Descriptions of AI Behavior. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3579612

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum.

Court of Justice of the European Union (2023). Judgment of 7 December 2023, *OQ v Land Hessen (SCHUFA Holding)*, Case C-634/21, ECLI:EU:C:2023:957.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus Actuarial Judgment. *Science, 243*, 1668-1674. https://doi.org/10.1126/science.2648573

Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the Jagged Technological Frontier: Field Experimental*

Evidence of the Effects of AI on Knowledge Worker Productivity and Quality (Harvard Business School Working Paper No. 24-013). https://ssrn.com/abstract=4573321

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*, 114-126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science, 64*, 1155-1170. https://doi.org/10.1287/mnsc.2016.2643

EU-OSHA (2025). *OSH Pulse 2025: Digitalisation* [Infographic]. European Agency for Safety and Health at Work. https://healthy-workplaces.osha.europa.eu/sites/hwc/files/hwc/infographic/OSH-pulse-2025-digitalisation_infographic_EN.pdf

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals, 14*, 627-660. https://doi.org/10.5465/annals.2018.0057

Goh, E., Gallo, R., Hom, J., Strong, E., Weng, Y., Kerman, H. et al. (2024). Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Network Open, 7*, e2440969. https://doi.org/10.1001/jamanetworkopen.2024.40969

Goh, E., Gallo, R., Strong, E. et al. (2025). GPT-4 Assistance for Improvement of Physician Performance on Patient Care Tasks: A Randomized Controlled Trial. *Nature Medicine*. https://doi.org/10.1038/s41591-024-03456-y

Green, B. (2022). The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Computer Law & Security Review, 45*, Article 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics, 6*, 107-128. https://doi.org/10.3102/10769986006002107

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine, 21*, 1539-1558. https://doi.org/10.1002/sim.1186

Kahneman, D., & Klein, G. (2009). Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist, 64*, 515-526. https://doi.org/10.1037/a0016755

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A Flaw in Human Judgment*. Little, Brown Spark.

Katz, D., Baptista, J., Azen, S. P., & Pike, M. C. (1978). Obtaining Confidence Intervals for the Risk Ratio in Cohort Studies. *Biometrics, 34*, 469-474. https://doi.org/10.2307/2530610

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. *Quarterly Journal of Economics, 133*, 237-293. https://doi.org/10.1093/qje/qjx032

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes, 151*, 90-103. https://doi.org/10.1016/j.obhdp.2018.12.005

McGraw, K. O., & Wong, S. P. (1992). A Common Language Effect Size Statistic. *Psychological Bulletin, 111*, 361-365. https://doi.org/10.1037/0033-2909.111.2.361

McKinsey & Company (2025, November 5). *The State of AI in 2025: Agents, Innovation, and Transformation* [McKinsey Global Survey]. https://www.mckinsey.com/featured-insights/artificial-intelligence/the-state-of-ai-2025

Meehl, P. E. (1954). *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors, 52*, 381-410. https://doi.org/10.1177/0018720810376055

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review, 46*, 192-210. https://doi.org/10.5465/amr.2018.0072

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 12 July 2024.

Regulation (EU) 2026/1744 (Digital Omnibus on AI), amending Regulation (EU) 2024/1689. *Official Journal of the European Union*, 24 July 2026.

Rubinstein, R. Y., & Kroese, D. P. (2017). *Simulation and the Monte Carlo Method* (3rd ed.). Wiley.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *Quarterly Journal of Economics, 69*, 99-118. https://doi.org/10.2307/1884852

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does Automation Bias Decision-Making? *International Journal of Human-Computer Studies, 51*, 991-1006. https://doi.org/10.1006/ijhc.1999.0252

Stevenson, M. T., & Doleac, J. L. (2019). *Algorithmic Risk Assessment in the Hands of Humans* (IZA Discussion Paper No. 12853). Institute of Labor Economics.

U.S. Census Bureau (2026, May 26). *AI Use at U.S. Businesses* [Census Bureau story]. https://census.gov/library/stories/2026/05/ai-use-businesses.html

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour, 8*, 2293-2303. https://doi.org/10.1038/s41562-024-02024-1

VanderWeele, T. J., & Ding, P. (2017). Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine, 167*, 268-274. https://doi.org/10.7326/M16-2607