ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

Leadership as an AI Governance Capability: A Contingency Model of Effective Human Oversight and Human, AI Complementarity

Imane Ouchen[1]; Ghizlane Chouay[2]; Amina Erryahi[3]; Zoheir Ahmed[4]

Author[1]; Co-Author[2;3;4]

[1]Laboratory of Interdisciplinary Laboratory for Economic, Econometric, and Managerial Research, Faculty of Legal, Economic, and Social Sciences, Mohammed FirstUniversity - Oujda, Morocco.
[2]Laboratory of Interdisciplinary Laboratory for Economic,Econometric, and Managerial Research, Faculty of Legal, Economic, and Social Sciences, Mohammed First University - Oujda, Morocco.
[3]Laboratory of Interdisciplinary Laboratory for Economic, Econometric, and Managerial Research, Faculty of Legal, Economic, and Social Sciences, Mohammed First University - Oujda, Morocco.
[4]Laboratory of Interdisciplinary Laboratory for Economic, Econometric, and Managerial Research, Faculty of Legal, Economic, and Social Sciences, Mohammed First University - Oujda, Morocco.

Publication Date: 2026/11/17

Abstract: Assigning a human reviewer to an AI-assisted decision does not ensure meaningful oversight. Reviewers also need the evidence, authority, and time to identify errors and secure appropriate corrections. Building on Ouchen, Chouay, and El Arabi's (2026) account of leadership as a governance capability, this article develops a contingency model of how those conditions can be established. The model links responsibility boundaries, decision orchestration, and collective learning to three mechanisms: calibrated reliance, actionable contestability, and verified learning feedback. Their expected effects depend on task evaluability, error detectability, review resources, and incentives. We distinguish improvement over human performance from synergy, defined as combined performance exceeding both standalone agents, and assess procedural safeguards separately. Five hypotheses guide a proposed research program that combines qualitative fieldwork, a team-randomized factorial experiment, and a subsequent learning intervention. Behavioral records and independent quality assessments provide the basis for testing the model, rather than favorable perceptions of leadership alone. The article contributes a testable explanation of when leadership practices may turn formal oversight into effective intervention. Human–AI complementarity is treated as an outcome to establish, not an assumed benefit of collaboration. No original empirical data are reported.

Keywords: *AI Governance; Leadership; Human Oversight; Human–AI Complementarity; Calibrated Reliance; Organizational Learning; Experimentation.*

How to Cite: Imane Ouchen; Ghizlane Chouay; Amina Erryahi; Zoheir Ahmed (2026) Leadership as an AI Governance Capability: A Contingency Model of Effective Human Oversight and Human, AI Complementarity.
International Journal of Innovative Science and Research Technology, 11(11), 464-478.
https://doi.org/10.38124/ijisrt/26nov456

I. INTRODUCTION

A manager may approve an AI recommendation without understanding the evidence behind it. An employee may notice an inconsistency yet lack the authority to pause a decision. A team may correct the same error repeatedly without revising the procedure that allows it to recur. These illustrative situations expose a gap between assigning human responsibility and making that responsibility actionable. Understanding this gap requires attention to the work through which oversight is exercised.

Ouchen, Chouay, and El Arabi (2026) describe leadership as a governance capability in organizations using AI. Their framework links leadership configurations to the interpretation of recommendations, their incorporation into organizational routines, and subsequent outcomes. The present article develops one part of that argument: how observable leadership practices might improve individual

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

decision episodes. We specify the mechanisms through which these practices are expected to operate and a research design for testing them.

The need for such an account follows from evidence that collaboration is not reliably superior to independent performance. Vaccaro, Almaatouq, and Malone's (2024) meta-analysis finds that human–AI combinations improve performance relative to humans alone on average, yet underperform the better standalone agent. Results also differ across tasks. Better performance by an assisted user therefore establishes augmentation, but does not establish that combining the two agents is the best-performing arrangement.

We ask how, and under what conditions, governance practices enacted by team leaders convert formal oversight into effective intervention and human–AI complementarity. Our argument centers on how people use recommendations, exercise intervention rights, and learn from evaluated decisions. Accepting sound advice and correcting flawed advice are both necessary concerns; neither frequent acceptance nor frequent rejection is inherently desirable.

The contribution is a model that separates governance practices from the behaviors they are intended to influence and the outcomes by which they should be judged. We identify conditions that may weaken or reverse the expected relationships and propose experimental comparisons that distinguish quality gains from synergy and procedural protection. This approach makes the leadership explanation open to disconfirmation, including evidence that technical or resource differences account for the gains instead.

II. THEORETICAL POSITIONING AND APPROACH

➢ *Governance Practices and the Unit of Analysis*

The original framework links leadership and AI governance to a broad range of organizational and socioeconomic consequences. We retain its premise that work arrangements shape the use of recommendations, while limiting the present explanation to teams and decision episodes. Claims about employment, wages, or inequality require additional evidence about how local changes accumulate across organizations and labor markets.

We organize the analysis around three sets of practices. Responsibility boundaries specify authorized uses, intervention rights, and escalation thresholds. Decision orchestration determines who handles a case, when each contribution is made, and where verification enters the workflow. Collective learning concerns how teams review errors and revise their routines. These are governance functions that may coexist within an organization, rather than mutually exclusive leadership styles.

Orlikowski's (2000) practice lens provides a useful foundation for this approach by locating technology use in situated activity. Decision outcomes depend on interfaces, model quality, professional expertise, and resources as well as managerial intentions. We therefore understand leadership as the work of establishing and sustaining governance practices, including resolving disputes about their application within an organizational unit.

➢ *Conceptual Design and Scope*

Following Jaakkola's (2020) distinction between conceptual research designs, this article develops a model through theory synthesis. Leadership-as-governance provides the organizational starting point; reliance research supplies the behavioral distinction between trust and use; oversight scholarship explains why responsibility may lack practical force; and learning research supplies the temporal mechanism. These literatures are selected for these explanatory roles, not treated as interchangeable evidence that leadership is effective. The procedure is an analytical synthesis rather than a systematic search, and no exhaustive coverage or screening counts are claimed.

The analysis uses three questions to evaluate each proposed relationship. First, what observable practice changes? Second, what behavioral response could connect that change to the outcome? Third, what condition could prevent that response? We retain a construct only when it has a role distinct from an outcome or a presumed precondition. This procedure does not validate the resulting model, but makes its derivation and points of potential failure explicit.

The model applies primarily to systems that recommend, classify, predict, or propose actions in workflows where people can still intervene. Systems that execute autonomous actions also require analysis of technical shutdown and recovery arrangements. The decision episode is the unit of observation. Governance interventions operate at the team level, which is also the proposed unit of randomization.

➢ *Positioning Relative to Adjacent Explanations*

Raisch and Krakowski (2021) explain the interdependence of automation and augmentation in management. Their paradox perspective supports examining work arrangements over time. The narrower question here concerns how a particular recommendation becomes subject to warranted correction. Lee and See (2004) address appropriate reliance, while the present model asks how review resources and intervention rights allow a suitable judgment to become an organizational action. The contribution is therefore an organizational extension of behavioral reliance, not a replacement for it.

Institutional oversight scholarship already recognizes the fallibility of human reviewers. Laux (2024) develops institutional principles that anticipate that fallibility, making competence and accountability relevant to the design of oversight. Kellogg, Valentine, and Christin (2020) show why algorithmic control must also be understood in relation to managerial power and worker responses. We use these perspectives to explain why an employee can identify a defect yet remain unable or unwilling to act. Such an

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

episode cannot be explained adequately through a trust rating alone.

Gonzalez et al. (2026) place reasoning, memory, attention, coordination, and governance within a collective-intelligence account of human-AI complementarity. Governance is consequently not an unoccupied theoretical space. Our more restricted contribution is to connect team-level practices to the conversion of a warranted concern into a timely authorized response and to specify experimental contrasts for testing that conversion. The model does not claim priority for integrating human and organizational factors in AI teaming.

III. DISTINCTIONS REQUIRED FOR RIGOROUS EVALUATION

➢ *Formal and Effective Oversight*

Formal oversight is the documented assignment of review or approval responsibility. Effective oversight is demonstrated when a process identifies a material defect, selects an appropriate response, and implements it within the time available. Relevant defects include factual errors, unsuitable assumptions, unauthorized actions, and failures against domain-specific criteria.

Oversight can fail in different directions. It becomes symbolic when responsibility is assigned without the means to exercise it. It becomes excessive when reviewers alter correct recommendations or add checks whose benefits do not justify their burden. Acceptance and rejection rates alone cannot distinguish these situations. What matters is whether intervention was warranted and whether it improved the handling of the case.

Green (2022) questions policies that treat the requirement for human oversight as a safeguard in itself. NIST's (2023) risk management framework also recognizes the need to differentiate roles and responsibilities. The issue is therefore how governance requirements work in practice, including whether formal procedures give reviewers the resources and authority to act.

➢ *Human Augmentation and Synergy*

Let QH represent mean quality for humans working alone, QA mean quality for AI alone, and QC mean quality for the combined configuration. Comparisons use comparable tasks and a common criterion on which higher values indicate better performance. We distinguish two quantities:

Human augmentation = QC – QH

Human-AI synergy = QC - max(QH, QA)

These measures can point in different directions. In a hypothetical example, suppose humans score 70, AI scores 85, and the combined configuration scores 80. Human augmentation is then 10 points, while synergy is -5 points. AI improves the user's performance, but combining the agents still yields a lower score than AI alone.

Performance comparisons inform deployment without determining it. Professional or institutional requirements may rule out autonomous decision-making, in which case AI alone serves as an experimental benchmark outside operational use. Resource differences also matter. Superior quality achieved with more time, information, or expenditure does not establish superior efficiency; those inputs must be reported alongside the outcome.

➢ *Immediate Quality and Judgment over Time*

Decision quality and procedural protection require separate evidence. Human involvement may be necessary to provide recourse or identifiable accountability even when it does not increase average accuracy. These functions should be assessed directly. Faster processing cannot be assumed to compensate for failing to meet an applicable requirement.

Immediate performance also needs to be separated from the development of unaided judgment. Assistance may promote learning, leave skills unchanged, or reduce opportunities to exercise them. Brynjolfsson, Li, and Raymond (2025) report heterogeneous effects across workers and evidence of learning in their study setting. Their findings do not support a general presumption of either deskilling or skill retention. Organizations that rely on experts to handle exceptions or operate during system outages need direct evidence of whether those capabilities are maintained.

IV. A GOVERNED COMPLEMENTARITY MODEL

➢ *Definition and Dimensions*

We use governed complementarity to describe an explicit arrangement for allocating tasks between humans and AI, verifying recommendations, and learning from decisions. The term identifies an architecture intended to enable complementarity. It does not imply that the arrangement has achieved synergy, which remains a comparative outcome requiring evidence.

Each set of practices addresses a different organizational problem. Responsibility boundaries clarify permissible uses and intervention rights. Orchestration aligns case allocation and the sequence of contributions with assessed capabilities. Collective learning turns validated observations into revised practices. Strength in one area does not remove weaknesses elsewhere: an incident review cannot repair an allocation rule unless its findings change that rule, and permission to challenge provides little help to someone unable to recognize an error.

The model proposes three intervening mechanisms. Calibrated reliance concerns selective use of recommendations; actionable contestability concerns the movement from justified disagreement to authorized action; verified learning feedback connects earlier decisions to subsequent knowledge and routines. Practices may influence more than one mechanism. An escalation rule, for example, may help people challenge an output while clarifying when routine reliance is inappropriate.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

➢ *Calibrated Reliance*

Calibrated reliance means using a recommendation in a way that matches its suitability for the decision at hand. Lee and See (2004) distinguish trust as an attitude from reliance as behavior. Someone who expresses distrust may still follow every recommendation under time pressure. Another person may trust the system within a validated domain while refusing an unsupported use. Trust ratings alone cannot establish whether reliance is appropriate.

Leaders may support calibration by communicating evaluated task boundaries, making relevant evidence available, and preserving independent judgment where it adds information. Explanations generated by AI do not necessarily provide that independence. In Bansal et al.'s (2021) experiments, explanations increased acceptance regardless of recommendation accuracy without further improving complementary performance. A persuasive rationale should therefore be evaluated rather than treated as evidence of reliability.

The value of review also depends on how human and AI errors overlap. When both agents fail on the same cases for similar reasons, adding a reviewer may contribute little independent information. Fügener et al. (2022) show that delegation arrangements and knowledge of relative capabilities affect joint performance. This supports evaluating allocation rules on cases that were not used to develop them.

➢ *Actionable Contestability*

Actionable contestability is the practical ability to report a defect, reach someone qualified to assess it, and obtain a timely response. It requires an escalation pathway, usable review resources, and authority to intervene. A suspension button offers limited protection if employees are penalized for using it. Encouraging discussion is similarly insufficient when no one can halt implementation.

Incentives shape whether these rights are exercised. If throughput is the only rewarded outcome, a justified review may impose a personal cost on the employee. If any acceptance preceding an incident is punished, employees may instead reject recommendations defensively. The relevant objective is warranted decision-making and reporting, assessed independently of the sheer number of approvals or challenges.

Psychological safety may help people express uncertainty. Edmondson (1999) associates it with learning behavior in the teams she studied, but its specific role in challenging AI remains a proposition of this model. Perceived freedom to speak up, authority to intervene, and the correctness of an intervention must be measured separately. Affected individuals may also need recourse beyond the team, requiring links to institutional governance.

➢ *Verified Learning Feedback*

Verified learning feedback connects past decisions to changes in future practice. Teams need to record why an intervention occurred, assess whether it was justified, identify recurring failures, and revise training or delegation boundaries accordingly. A record of disagreement is not enough. Without checking its validity, an organization may institutionalize a human mistake.

Team leaders can support this process by making qualified staff available for review and following through on agreed changes. Over time, comparable errors should become less frequent or easier to detect if the mechanism works as proposed. Delayed consequences and defensible differences of interpretation complicate that assessment. Feedback should preserve such uncertainty; a supervisor's agreement cannot substitute for a justified evaluation criterion.

➢ *Boundary Conditions*

Task evaluability concerns whether decision quality can be assessed against defensible criteria. Error detectability concerns what a reviewer can recognize from the evidence available at the time of intervention. A decision may be assessable after six months yet difficult to verify before implementation. Eventual outcome information therefore does not establish the feasibility of effective prior review.

Review resources include expertise, protected time, accessible evidence, and a manageable workload. Incentives are the rewards and penalties people encounter in practice. These conditions must be measured independently of decision outcomes. Otherwise, the model risks treating the presumed causes of effective governance as evidence that governance is already effective.

The severity and reversibility of a decision help determine how much oversight is warranted, but do not guarantee that more review will work. Serious, irreversible decisions may also contain errors that reviewers cannot detect. Adding approval signatures may then offer little protection. Restricting use or changing the technical architecture must remain possible responses, because leadership practices cannot compensate for every limitation of the tool.

➢ *The Conversion of Evaluative Judgment into Authorized Action*

The model's organizing claim concerns a conversion problem. A reviewer may know that a recommendation is unsuitable without being able to change what the organization does. Responsibility boundaries are expected to alter this transition by specifying who can pause, escalate, and resolve a case. Their incremental benefit should be greatest where employees possess relevant evaluative competence but previously lacked a usable intervention route. Where that route is already effective, the same formal intervention may add little. This prediction distinguishes the account from one in which information or skill alone determines oversight.

Orchestration addresses a different constraint: competent intervention must be applied to cases where it can contribute. Routing can improve outcomes when it exploits

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

differences in human and AI capabilities estimated independently of the evaluation cases. Yet routing may fail if reviewers share the system's errors or cannot verify the cases assigned to them. The positive interaction predicted by H3 is therefore conditional rather than logically necessary. An efficient allocation rule may work without an additional responsibility package, and redundant controls may even reduce its advantage.

Verified feedback closes the process across decision episodes. Its expected benefit depends on whether the organization can distinguish warranted interventions from mistaken ones and enact the resulting changes. Feedback volume alone has no favorable prediction. This temporal argument links the decision episode to team routines without assuming that individual learning and organizational adaptation are equivalent outcomes. The experiment below tests selected implications of this account; it does not test every link in the dynamic model.

➢ *Research Hypotheses*

The five hypotheses below translate the model into tests of specified interventions and observable outcomes. They allow benefits on one criterion to coexist with costs on another.

- H1-Responsibility and calibrated reliance. Holding orchestration constant, an enhanced responsibility and contestability intervention reduces adherence to incorrect recommendations when users have the information needed to verify them. Rejection of correct recommendations will be assessed separately to determine whether the intervention instead encourages indiscriminate distrust.
- H2 - Contingent orchestration. Holding the responsibility intervention constant, an allocation rule based on relative performance estimated on a development set improves decision quality on an independent evaluation set compared with a uniform usage rule. Such a gain does not, by itself, establish superiority to both standalone agents.
- H3 - Complementarity between interventions. Contingent orchestration produces a larger quality gain when the enhanced responsibility intervention is present. This prediction concerns an interaction contrast on the prespecified outcome scale. Positive main effects for both interventions would not be sufficient to support it.
- H4 - Review conditions. The responsibility intervention reduces erroneous adherence more strongly when baseline error-detection competence is higher and case verification costs are lower. These characteristics will be measured before treatment or defined independently of the behaviors it produces. Extensions of the protocol will examine evaluability and incentives as additional boundary conditions.
- H5 - Learning and autonomy. Among teams with comparable prior AI exposure, a structured error review grounded in independent assessment improves performance on a delayed unaided test relative to a control activity of equal duration. The test measures retention or transfer over the specified interval, rather than indefinite preservation of competence.

An adverse effect would challenge the corresponding prediction, as would an estimate precise enough to rule out the smallest effect of interest. Statistical nonsignificance alone would not establish the absence of an effect. Interpretation will rely on effect sizes and uncertainty intervals to distinguish evidence against a prediction from an inconclusive test.

➢ *Measurement Strategy*

- *Separating Practices from Outcomes*

Measurement will draw on organizational documents, observations of routines, interaction logs, and competence tests. Interviews and questionnaires will help explain participants' interpretations, but will not supply the only measures of both governance and effectiveness. Independent quality assessors will be blinded to experimental assignment.

Table 1 Proposed Measures of Practices and Mechanisms

Construct	Primary Measure	Measurement Consideration
Governance practices	Observed rules, resources, and rights	Distinguish availability from implementation
Calibrated reliance	Erroneous adherence and unwarranted rejection on diagnostic cases	Report each denominator
Actionable contestability	Defects reported, adjudicated, and corrected in time	Assess justification independently
Verified feedback	Assessment-based reviews and documented changes	Distinguish activity from learning
Quality and costs	Independent score, time, and total cost	Do not aggregate without an explicit rule
Unaided competence	Delayed test without AI on new cases	Assess retention and transfer

Governance dimensions will first be examined separately. A formative resource inventory need not contain strongly correlated components, since each component may contribute a different resource. A reflective perceptual subscale, by contrast, requires evidence about its structure, reliability, and discriminant validity. Including a score or weighting scheme in the protocol does not validate it.

- *Observing Disagreements and Corrections*

Disagreements between an initial human judgment and an AI recommendation provide useful diagnostic cases. If

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

the human is initially correct and AI is wrong, following AI produces an observable loss. If the human is wrong and AI is correct, rejecting its recommendation forgoes a potential improvement. We will report these rates separately, with their denominators and uncertainty, rather than collapse them into an agreement rate.

Cases in which both agents are wrong also matter. Verification or escalation may still produce a correction, while agreement may conceal a shared mistake. Open-ended tasks will be scored using a rubric and an adjudication procedure, with agreement among assessors examined. Final accuracy and the quality of the supporting reasoning will remain separate outcomes.

Initial judgments will be collected on diagnostic cases before recommendations become visible. The module will be the same across conditions and, where necessary, separate from workflows used to test presentation order. Because expecting later assistance may change initial effort, these judgments will not be used as the principal human-only benchmark.

V. PROPOSED EMPIRICAL RESEARCH PROGRAM

➢ *A Preparatory Qualitative Phase*

The preparatory phase will examine teams whose tasks and oversight arrangements differ. Case selection will seek variation in evaluability, review resources, and stakes, including arrangements that appear unsuccessful. Interviews with leaders, users, and technical staff will be examined alongside documented procedures and observed decisions.

Fieldwork will trace how recommendations reach implementation: who receives the evidence, where defects are noticed, who can act, and how reports are handled. These observations will refine intervention content and identify difficulties in understanding the proposed measures. They will not establish causal effects or justify selecting only routines associated with favorable outcomes.

➢ *Tasks and Outcome Definition*

An initial experiment should use a bounded domain, such as request triage or simulated case analysis. Deliberately introduced errors must not influence real decisions affecting individuals. A development set will support estimates of relative capabilities and the design of allocation rules. Evaluation will use an independent set.

To make the design concrete, a reference implementation could use simulated administrative eligibility screening. Each case would contain a fictional application and a fixed set of eligibility rules; the required response would be a classification and identification of the evidence supporting it. No real applicant would be evaluated. This is a proposed implementation for methodological illustration, not a recruited field site or a claim about the generalizability of administrative screening to all organizational decisions.

The primary outcome would be the proportion of cases correctly classified on the independent evaluation set. A prespecified secondary outcome would distinguish erroneous approvals from erroneous denials; supporting-evidence quality, escalation, time, and total review cost would be reported separately. A failure to resolve a case within the task window would be coded under a rule fixed before collection, rather than silently removed. A pilot would establish whether the rules produce a sufficiently reliable reference answer and usable variation in human and AI performance.

The study will retain instructions, model versions, parameters, documentary resources, and access budgets. Cases may vary in difficulty and include plausible errors. If errors are oversampled to make review behavior easier to study, raw findings will be distinguished from estimates weighted to a documented operational distribution. Their experimental frequency will not be interpreted as their natural prevalence.

➢ *A Team Randomized Factorial Experiment*

The main experiment crosses two factors: R, an enhanced responsibility and contestability intervention, and O, orchestration based on relative performance. Randomization will occur at the team level, with proposed stratification by organization and baseline performance. A large number of decisions within a few teams cannot substitute for a sufficient number of independently assigned teams.

Table 2 Conditions in the Main Factorial Design

Condition	Responsibility and Contestability	Orchestration
R0O0	Common foundation	Uniform procedure
R1O0	Enhanced intervention	Uniform procedure
R0O1	Common foundation	Prespecified contingent allocation
R1O1	Enhanced intervention	Prespecified contingent allocation

Every condition will receive minimum training, usage boundaries, data protection, and necessary avenues for recourse. R1 will add explicit operational responsibility, a designated person for escalation, appropriate suspension rights, and a review budget. O1 will use an allocation rule fixed before evaluation, while O0 will follow a clearly specified uniform procedure. All conditions will use the same interface and general training content; attention from study staff will be matched where possible.

R1 will exclude structured learning reviews during the main phase so that a separate intervention can test H5 afterward. Standardized, neutral decision recording will continue in all conditions for measurement purposes. Any

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

extra review time provided by R1 will be counted as part of the treatment's resource requirement, rather than assumed to be held constant.

This design estimates the effects of bundles of practices. It cannot isolate protection for contestability as the source of an R1 effect, nor can it establish the effect of a general latent leadership trait. Documentation of leader involvement and implementation fidelity will show what was delivered and how consistently teams used it.

➢ *Standalone Benchmarks*

Separate assessments will measure professional performance without AI and AI performance without human intervention on cases drawn from the same distribution. Independent groups or counterbalanced parallel sets will limit memorization among professionals. Assisted participants' initial judgments will not be the sole basis for reconstructing human-only performance.

The standalone AI procedure will be documented and repeated if output variability warrants it. Available information, time budgets, and costs will be reported for each configuration. Analyses will test QC - QH and QC - QA separately. Positive synergy will require superiority to both benchmarks under prespecified thresholds and an inference rule. An advantage over another assisted experimental group cannot establish that conclusion.

➢ *A Separate Learning Module*

After the main phase, teams will be randomized again to a structured error review or a control activity of equal duration, balancing assignment across the R and O conditions. Review sessions will examine human and AI decisions using independent criteria. Control sessions will not include that systematic comparison.

A delayed test on new cases, completed without AI, will assess unaided competence and transfer. Its timing will be specified before data collection to suit the domain. Controlled exercises involving degraded or withdrawn AI support could provide additional evidence about recovery capacity, but would remain distinct from the primary test of H5. The study will track use of external tools, attrition, and exchanges between teams.

➢ *Statistical Power and Analysis*

Simulation-based power calculations will incorporate the smallest effect of interest, the number of teams, intraclass correlation, case variability, and attrition. Behavioral analyses will also depend on how often humans and AI disagree. Power for the R × O interaction will be assessed separately, since it may require more teams than the main effects. Sample size will not be justified through a generic respondent-count rule.

The primary analysis will estimate intention-to-treat effects. For a continuous quality outcome, a possible model is:

$$Q_{sijt} = \beta_0 + \beta_R R_j + \beta_O O_j + \beta_{RO} R_j O_j + \gamma' X_{ij} + \lambda_t + u_j + v_{ij} + a_s + \varepsilon_{sijt}$$

The indices s, i, j, and t denote case, individual, team, and period. X contains prespecified pretreatment covariates, and λ_t denotes period effects. The terms u_j, v_ij, and a_s represent team-, individual-, and case-level random intercepts; ε_sijt is the episode-level residual. Individuals are nested within teams, whereas cases may be crossed with teams. The model illustrates the dependence structure rather than prescribing a distribution for every outcome. Binary accuracy will be analyzed with an appropriate prespecified model, and effects will also be reported on the probability scale.

Inference will follow the team-level assignment scheme. The analysis plan will specify methods suited to the number of clusters or randomization inference consistent with the design. Decision episodes will not be treated as independent assignments. Primary outcomes, interaction contrasts, missing-data procedures, and multiplicity adjustments will be set before access to the results.

Synergy analyses will use the two contrasts with standalone performance, accompanied by simultaneous uncertainty intervals. This avoids selecting the better agent on the evaluation data and then treating its observed score as a benchmark without uncertainty. Time, cost, and error severity will be reported separately from the primary quality outcome.

➢ *Testing Mechanisms without Overstating Causality*

Randomizing R and O identifies their total effects in the study setting. It does not randomize reliance, review effort, or psychological safety. Associations between these variables and quality may be consistent with the proposed mechanisms, but cannot alone establish them causally.

Mediation analyses will be secondary, with explicit identification assumptions and sensitivity analyses addressing the concerns discussed by Imai, Keele, and Tingley (2010). Posttreatment implementation measures will not automatically enter the primary model as controls. Isolating components such as incentives or intervention rights will require additional experiments.

➢ *Testing Mechanisms without Overstating Causality*

- *Distinguishing Intervention Effects from Mechanism Claims*

Randomization of R and O supports identification of their intention-to-treat effects under the assignment and follow-up conditions specified in the protocol. It does not randomize calibrated reliance, review effort, or psychological safety. A team with greater review effort may also have stronger expertise or a more supportive supervisor, each of which could influence quality. Evidence that

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

assignment changes a proposed mediator, and that the mediator predicts the outcome, is therefore consistent with the model without establishing that the mediator transmits the intervention effect. The factorial design identifies contrasts between governance arrangements; the causal explanation of those contrasts requires additional assumptions or experimental variation.

Each mediation analysis will specify the treatment contrast, mediator, outcome, and observation window before estimation. For example, the question may concern the R1 versus R0 contrast at a fixed level of O, with warranted escalation measured before final adjudication and independently assessed quality as the outcome. This differs from asking whether escalation explains an effect averaged across both orchestration conditions. Following Imai, Keele, and Tingley (2010), direct and indirect effects will be defined independently of the regression model used to estimate them. Treatment–mediator interactions will be allowed when theoretically justified; a single product of regression coefficients will not be assumed to represent every mediation effect.

- *Identification Assumptions and their Limits*

Random assignment addresses confounding of the assigned intervention, but ordinarily does not remove confounding between a mediator and the outcome. Identification of a natural indirect effect requires assumptions concerning unmeasured mediator–outcome confounding, treatment-induced confounding, consistency, and adequate support for the comparisons. Imai, Keele, and Yamamoto (2010) show why mediation estimates and sensitivity analyses must be interpreted in relation to their identifying assumptions. In this setting, relevant pretreatment common causes include domain competence, prior AI experience, workload, and task difficulty. Measuring them improves the plausibility of adjustment without proving that all relevant confounding has been removed.

Review effort illustrates a difficulty that cannot be resolved by adding every available variable to a regression. R1 may increase effort, which may affect both escalation and decision quality. In an analysis treating escalation as the mediator, effort is then a possible treatment-induced mediator–outcome confounder. Conditioning on it mechanically can remove part of the effect under study or create selection bias, depending on the causal structure. Similar problems arise if analysis is restricted to workers who comply fully with the procedure or cases that reach escalation. The primary intention-to-treat comparison will therefore remain separate from mediation and implementation analyses. Where the assumptions required for a natural-effect decomposition are implausible, that decomposition will not be given a causal interpretation.

Measurement must establish a plausible sequence rather than collect all variables after the decision. Baseline competence and prior experience will be assessed before assignment. Review actions and escalation will be time-stamped before the final outcome is scored. Psychological safety, if examined as an intervention-responsive mediator, will be measured before the later decisions it is intended to explain; a baseline measure may instead serve as a moderator. Verified learning feedback will be linked to subsequent decisions or delayed competence, not used to explain the same decision that generated it. Temporal ordering reduces ambiguity but does not eliminate unmeasured confounding.

- *Avoiding a Mediator that Restates the Outcome*

Some reliance measures are close to the quality criterion by construction. When an incorrect AI recommendation is accepted and the final answer is wrong for that reason, erroneous adherence and error are mechanically linked. Treating their association as a discovered causal mechanism would overstate the evidence. Behavioral measures will therefore distinguish observable actions, such as evidence consultation, an authorized pause, and a completed escalation, from independently evaluated outcomes. Diagnostic reliance rates remain useful for describing where performance gains or losses occur, but their overlap with accuracy will be stated explicitly. Correct final answers reached through invalid reasoning will also be distinguished from justified correction where the rubric permits it.

The proposed mechanisms can influence one another. A protected escalation route may increase verification, while successful review may change future reliance. Separate single-mediator models will not be summed as if they decomposed the total effect into independent pathways. A proposed ordering will be justified before fitting a sequential model, and competing orderings will be acknowledged where the design cannot distinguish them. Team-level practices, individual behavior, and case-level outcomes also require compatible measurement and uncertainty estimation. Repeated cases provide additional observations, but do not increase the number of independently randomized teams. Between-team associations will not automatically be interpreted as within-person mechanisms.

- *Sensitivity Analysis and Evidential Thresholds*

For any estimated indirect effect, the report will state the identifying assumptions, model specification, uncertainty interval, and sensitivity to plausible unmeasured mediator–outcome confounding. Sensitivity parameters will be interpreted using domain-relevant scenarios or comparisons with measured confounders where feasible. Alternative operationalizations and uncertain adjudications will be examined without presenting a favorable specification as uniquely authoritative. A significant indirect effect will not certify a mechanism, and a nonsignificant total effect will not by itself exclude offsetting pathways. Reports will avoid an unqualified proportion-mediated statistic when the total effect is small, imprecise, or opposite in sign to an indirect effect.

A follow-up experiment could hold review time and information access constant while randomizing a usable pause-and-escalation right. Its effect on independently verified correction would test a specific implication of the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

authority account more directly than the bundled R intervention. Another experiment could vary performance information while retaining the same escalation structure to examine an information-based explanation. These manipulations would identify the effects of the assigned components, not necessarily of the observed mediator itself: a pause right may change attention as well as authority. Claims that a component isolates one pathway would therefore require evidence that competing pathways have not also changed.

- *Integrating Evidence without Overstating the Conclusion*

The strongest feasible interpretation will combine randomized intervention effects, temporally ordered process records, independent outcome assessment, and component experiments whose predictions differ across explanations. Qualitative observations can show why a concern failed to reach an authorized reviewer, while logs can establish whether the predicted sequence occurred. Neither source substitutes for identification assumptions. Findings will be described as intervention effects, evidence consistent with a mechanism, or causally identified mediation according to the design and assumptions actually supported. This distinction keeps the explanatory ambition of the leadership model proportional to the evidence available.

➢ *Estimands and Selection Bias*

Let μ_{ro} denote the mean outcome under assigned responsibility condition r and orchestration condition o. The simple R effect at O = 0 is $\mu_{10} - \mu_{00}$; the simple O effect at R = 0 is $\mu_{01} - \mu_{00}$. The interaction is $\mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$. Marginal effects average over the other factor's randomized levels. In the linear model above, βR and βO are simple effects at the other factor's reference level. Both simple and marginal contrasts will be defined in the analysis plan, with H3 evaluated through the interaction rather than a comparison of separate significance tests.

H1 will use a common diagnostic module with identical, prespecified AI recommendations across assigned conditions. Its primary behavioral endpoint is the proportion of prespecified AI-error cases on which the final response retains the erroneous recommendation, defined before treatment assignment. Rejection of correct recommendations is a separate outcome. Disagreement-conditioned rates are diagnostic: if treatment changes initial judgment, conditioning on disagreement selects different participants or cases across groups and does not preserve the randomized comparison. Those rates will therefore not be interpreted as causal effects without additional identification assumptions.

The enhanced responsibility package will specify an escalation recipient, a usable pause procedure, and the review time permitted for those cases. The orchestration package will apply a development-set allocation rule based only on information available before the decision. Where O1 allocates a case directly to AI, its outcome remains part of the assigned configuration's performance even if no user reviews it. Person-level behavioral measures will therefore be assessed in the common diagnostic module, rather than treating the selectively human-reviewed O1 cases as comparable to all O0 cases.

The learning module will estimate the effect of assignment to structured review on delayed unaided performance, adjusting for prespecified baseline competence and the earlier R/O assignments. A claim about recovery during an outage will require a separate recovery outcome. The final number of teams, case count, test interval, and smallest effect of interest remain pilot-dependent design parameters; specifying them arbitrarily would create apparent precision without an empirical basis. Cluster recruitment, allocation, and losses will be documented using the reporting logic discussed by Campbell et al. (2012), adapted to the organizational setting.

➢ *Intervention Development and Implementation Fidelity*

The qualitative phase will inform an intervention manual that specifies what each condition receives and what team leaders are expected to do. For R1, the manual will distinguish the right to raise a concern from the authority to pause implementation and the responsibility to resolve the concern. It will identify the escalation recipient, the evidence required for review, and the time window for responding. These specifications will be adapted to the domain before randomization and then held stable during the confirmatory phase. Adaptations made after assignment will be documented as departures from the planned intervention.

For O1, the allocation manual will define the case characteristics used to route work and explain how those characteristics can be assessed before the final outcome is known. Allocation must not rely on information that becomes available only after the decision. The development set will be used to estimate the relative performance of humans and AI within sufficiently supported task categories. Categories with too little evidence will follow a prespecified fallback rule, such as escalation or assignment to a qualified reviewer. The evaluation set will remain inaccessible when these choices are made.

Implementation fidelity will be assessed through process indicators collected independently of decision quality. Relevant indicators include whether the designated escalation actor was available, whether review time was provided, and whether the routing rule was followed. A fidelity checklist will distinguish an intervention that was delivered but ineffective from one that was never delivered as intended. These observations will explain implementation and support secondary analyses; they will not replace the intention-to-treat estimate or justify excluding inconvenient teams after results are known.

➢ *Outcome Adjudication and Decision Traceability*

Before data collection, domain experts will define the primary quality criterion and identify defects considered material. For tasks with a verifiable answer, the reference answer and the evidence supporting it will be established independently of participants' responses. For open-ended tasks, a scoring rubric will distinguish relevant dimensions,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

such as factual accuracy, completeness, and consistency with the stated decision criteria. The primary outcome will be selected in advance; secondary dimensions will not be substituted because they produce more favorable findings.

Assessors will receive standardized instructions and practice cases outside the confirmatory set. They will score independently before adjudication, and agreement before reconciliation will be reported. Persistent disagreements will be retained as evidence of evaluation uncertainty rather than concealed by a final consensus score. If a case cannot be adjudicated reliably under the rubric, its handling will follow a prespecified rule. Exclusion decisions will be made without access to treatment-specific effects wherever feasible.

A decision trace will link the case presented, the AI output available to the participant, the final response, and any review or escalation. Where the diagnostic module collects an initial unaided judgment, that response will be stored separately from the final decision. Timestamps will distinguish time spent on the case from time waiting for an escalation response. The same recording requirements will apply across conditions so that additional measurement does not inadvertently become a treatment available to only one group.

Traceability also allows the analysis to distinguish noticing a defect from successfully resolving it. A participant may correctly identify a problem but fail to obtain timely action; another may modify the output for an invalid reason and nevertheless reach the correct answer by chance. Both patterns would be obscured by final accuracy alone. Process evidence will therefore be used to interpret the outcome, while preserving the distinction between explanatory observations and causal tests of mediators.

- *Missing Data Contamination and Protocol Deviations*

Missing observations may arise at several levels: an individual can skip a case, a worker can leave the study, or an entire team can withdraw. The study will report the number and reasons at each level and compare missingness across assigned conditions. A participant's refusal to finalize a decision will not automatically be coded as missing. If abstention or escalation is a permitted response, it must be represented in the outcome framework, together with its consequences for turnaround time and resolution.

The analysis plan will distinguish technical data loss from unresolved decisions and delayed outcomes. The choice of imputation or another missing-data method will depend on the outcome, the study structure, and stated assumptions about missingness. Sensitivity analyses will examine plausible departures from those assumptions. A complete-case analysis may be reported, but it will not be treated as sufficient merely because it is easy to implement. No method can recover information lost through withdrawal without assumptions that should be made visible.

Contamination is possible when teams share leaders, training materials, or informal advice. Recruitment and scheduling will seek to reduce these exchanges without imposing unrealistic separation on ordinary work. Shared supervision and cross-team communication will be mapped before assignment. Where substantial spillovers are unavoidable, the randomization unit may need to move to a higher organizational level, with corresponding changes to the power calculation. Team-level assignment is a design choice to justify, not a requirement to preserve regardless of the setting.

Protocol deviations will be logged with their timing, cause, and likely scope. Model outages, unplanned updates, access to external tools, and changes in reviewer availability may alter the intervention actually received. Analyses addressing such deviations will remain secondary to the assigned-condition comparison. If an event makes the intended contrast uninterpretable, the report should say so directly rather than redefine the treatment after seeing the outcomes.

- *Heterogeneity Robustness and the Interpretation of Null Findings*

The primary analysis will estimate average intervention effects, but averages can conceal meaningful variation. A limited set of moderators will be prespecified on theoretical grounds, including baseline error-detection competence and independently determined verification cost. Analyses will estimate interactions directly rather than infer moderation from significance in one subgroup and nonsignificance in another. Estimates for small subgroups will be presented with appropriate uncertainty and will not support broad claims about entire occupations or demographic groups.

Several robustness checks can address features of the decision tasks. Results may be examined with alternative defensible scoring rules, with and without cases whose adjudication remains uncertain, and under documented operational weights when errors have been oversampled. Such analyses will be labeled according to whether they were prespecified. A favorable result that depends on a single scoring choice should lead to a narrower interpretation, even when the preferred specification is statistically significant.

Null findings require particular care. Wide intervals may indicate that the study cannot distinguish a useful effect from a harmful one. Narrow intervals that exclude the smallest effect of interest provide more direct evidence against the predicted practical benefit. Equivalence or noninferiority claims, if pursued, will require margins and tests specified in advance. A small p value is not the sole criterion for theoretical support, and a large p value is not a sufficient basis for declaring the interventions interchangeable.

- *System Versioning and Replication*

The study will archive the model identifier, observation dates, instructions, settings, permitted tools, and accessible information sources. Updates and outages will be logged so that changes in technical exposure are not

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

mistaken for intervention effects. If a stable version cannot be maintained, the analysis plan will specify how affected periods are handled. Replication with a newer system will be interpreted as a test under changed technical conditions, since relative capabilities and the suitability of a routing rule may have changed. The transparency statement below sets out access and confidentiality arrangements for research materials.

VI. DISCUSSION

- *Contribution to the Leadership as Governance Framework*

The model replaces a broad expectation that leadership improves AI use with a set of observable relationships. Its explanatory value depends on whether leadership practices change task allocation, the exercise of intervention rights, and learning from feedback. The contribution lies in connecting these practices to mechanisms and boundary conditions, then specifying comparisons that can test their effects. Trust, contestability, and learning themselves are established concerns.

Several competing explanations remain plausible. Better interfaces, stronger models, or more review time could account for the observed gains. If technical changes reproduce those gains without changes in managerial practice, the independent contribution attributed to leadership should be narrowed. Likewise, a successful intervention bundle would not establish that a hierarchical leader is always necessary. Professional or distributed governance may support some of the same mechanisms.

Adverse findings would also be informative. A more demanding review procedure may slow work without improving detection. An allocation rule may become unsuitable after a change in model capabilities. Feedback based on an incorrect assessment may entrench a flawed routine. Such results should lead to revisions of the explanation or intervention, rather than automatically be attributed to users' lack of motivation.

The conversion account also supplies a discriminating comparison for future work: an information-only intervention can be compared with an authority-and-escalation intervention under equal review time. If information alone produces the same gains, a lack of usable evidence may explain the result more parsimoniously. If an escalation route adds value primarily among competent reviewers, the authority-based account receives more specific support. Neither result would establish that a leader's personal style caused the change. Establishing that stronger claim would require a design that varies leadership enactment while holding the organizational package constant.

- *Organizational Tradeoffs*

Organizations can begin by following actual decision episodes through their review process. The relevant evidence concerns whether defects were detected, reached a qualified reviewer, and were corrected in time, as well as whether later reviews led to justified changes. This diagnostic work can locate failures in the process, although it does not replace an evaluation of the intervention's effects.

Dashboards should separate warranted acceptance, erroneous adherence, useful correction, and unwarranted rejection. Buçinca, Malaya, and Gajos (2021) show that interventions encouraging deliberation can reduce overreliance while receiving less favorable user ratings. Satisfaction is therefore an incomplete guide to decision quality. The time and expense of oversight should also remain visible when evaluating additional checks.

More reported incidents may indicate that employees have become better able to raise concerns. Treating that increase as immediate evidence of deterioration could discourage reporting. Comparing reports with independent audits and decision logs can help distinguish changes in the underlying failures from changes in their visibility.

- *What Different Findings Would Mean for the Model*

The strongest support for the proposed account would combine an intervention effect with process evidence consistent with the predicted behavior. For example, improved quality under R1 would be more informative if independently assessed records also showed fewer warranted concerns left unresolved. Even then, the main experiment would identify the bundled intervention's total effect. The process evidence would strengthen the explanation's plausibility without resolving every causal question about the intervening mechanism.

Other patterns would support only part of the argument. If R1 reduces adherence to erroneous recommendations but leaves average quality unchanged, the benefit may be offset by rejection of correct advice or by failures elsewhere in the workflow. If O1 improves quality without an R × O interaction, performance-based allocation may be useful while the predicted reinforcement between interventions is unsupported. A favorable main effect should not be used to rescue an interaction hypothesis that the data contradict.

An improvement over humans alone, combined with no superiority over AI alone, would establish augmentation rather than synergy. Such an arrangement could still be appropriate where human accountability or recourse is required. Its justification would then rest partly on those functions and their costs, not on a claim that the combined configuration is technically superior. This distinction permits a useful organizational recommendation without overstating the performance evidence.

The learning results could also diverge from immediate performance. Better assisted decisions followed by weaker unaided performance would raise questions about the longer-term dependence created by the workflow. Improved delayed performance without an immediate quality gain would instead suggest that the review intervention serves a developmental function. Either pattern

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

would require examining exposure, test validity, and uncertainty before attributing the result to learning or deskilling. The value of separating outcomes is precisely that one favorable result does not stand in for all others.

- *Leadership Authority and the Distribution of Accountability*

Treating leadership as a governance capability does not imply that a team leader controls every condition required for effective oversight. Procurement determines which systems are available; technical teams influence interfaces and access to evidence; senior management allocates staffing and time; professional or institutional bodies may establish decision rights. The team leader operates within this distribution of authority. A theory of leadership should identify what leaders can change rather than absorb every organizational resource into the leadership construct.

This matters for accountability. Naming an employee as the final decision-maker may concentrate responsibility on someone who cannot inspect the system or delay its use. Under the model, that arrangement is a candidate case of symbolic oversight, not evidence that a human safeguard has failed through individual negligence. Research should examine the authority attached to responsibility and the route through which unresolved concerns can move beyond the team.

Distributed governance may offer an alternative implementation of the same mechanisms. A professional review group, an independent quality function, or a jointly defined escalation process may provide expertise and protection that a single manager cannot. The present model leaves that possibility open. Comparing these arrangements would help determine whether the explanatory work is done by leadership enactment, by the institutional distribution of rights and resources, or by their interaction.

The implications for managerial evaluation are equally specific. Leaders should not be judged solely on adoption rates, throughput, or the absence of reported incidents. Those measures may reward passive acceptance or suppress reporting. A more informative evaluation would consider whether justified concerns receive competent review and whether the organization acts on verified recurring failures. These remain proposed evaluation principles, whose effects would need to be tested rather than assumed.

- *The Economics of Review and Selective Human Involvement*

Human review uses scarce attention. A policy requiring detailed examination of every recommendation may direct expertise away from the cases that most need it. The model therefore supports studying selective review, while refusing the assumption that the most complex cases are always those best handled by humans. Allocation depends on assessed relative capabilities, available evidence, and the consequences of error. Difficulty for the human and difficulty for the AI need not coincide.

A practical analysis should examine the marginal value of additional verification. An extra review step is useful when its expected correction benefit justifies its burden under the domain's constraints. That judgment is empirical and normative: the probability of catching an error can be estimated, while the acceptable treatment of severe or irreversible consequences requires decisions by authorized stakeholders. A numerical productivity gain cannot settle every such judgment.

Selective review also creates a learning problem. If workers encounter only exceptional cases while routine work is automated, their experience may become narrower and the remaining decisions harder to evaluate. Conversely, reviewing carefully selected cases could provide valuable practice. The direction cannot be inferred from the allocation rule alone. Tracking case exposure and unaided competence would allow future studies to examine whether efficient routing preserves the expertise on which oversight depends.

Organizations should therefore report the resources required to sustain the arrangement, not only the resources used during an individual decision. Maintaining reference cases, training escalation staff, adjudicating disagreements, and updating allocation rules all carry costs. These activities may be necessary for reliable operation, but omitting them can make a supervised configuration appear less expensive than it is. The appropriate comparison includes both decision-level burdens and the continuing infrastructure of review.

- *From Error Review to Organizational Adaptation*

The learning mechanism concerns changes in routines as well as changes in individual knowledge. A team may learn that a particular category should no longer be delegated, that a required source is unavailable, or that an escalation threshold is set too late. These are organizational adaptations. A higher test score among individual workers would not by itself demonstrate that such adaptations occurred, just as a revised procedure would not establish that workers acquired transferable competence.

Future longitudinal work could follow how verified incidents lead to proposed changes, approval, implementation, and later evaluation. This would reveal whether organizations close the loop between recognizing a failure and changing the conditions that produced it. The observation window must be long enough for those steps to occur, but should not assume that every change is an improvement. Revised routines need subsequent assessment on cases independent of the incident that prompted them.

Feedback may become self-confirming when the organization evaluates AI recommendations using judgments already influenced by the same system. Agreement can then be mistaken for independent confirmation. The model's requirement for verified feedback is intended to prevent this inference, although complete independence may be difficult in practice. Studies should document the provenance of reference assessments

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

and acknowledge where shared information or expertise makes independence partial.

Organizational adaptation also depends on who can authorize change. Team members may identify a recurring defect while the relevant model or workflow is controlled elsewhere. Escalation beyond the unit becomes part of the explanation in those cases. A local team's inability to repair the underlying system should not be concealed by repeated case-level correction, which may sustain short-term performance while leaving the recurring source of failure untouched.

- *Boundary Conditions and a Cumulative Research Agenda*

The first experiment should prioritize a setting with sufficiently clear evaluation criteria to establish whether the proposed contrasts are measurable. Later studies can extend the model to settings with delayed outcomes, contested criteria, or weaker intervention rights. These extensions should specify which mechanism is expected to change and why. Labeling a study as a different sector is less informative than identifying differences in evaluability, detectability, resources, or incentives.

Comparative studies could examine whether the same governance practices work differently under distinct incentive arrangements. A throughput-focused environment may discourage use of an otherwise well-designed escalation pathway. A quality-focused environment may still produce defensive behavior if employees expect blame after any adverse outcome. The relevant variable is the incentive structure workers encounter, including informal consequences, rather than the wording of a formal policy alone.

A further priority is the stability of relative capabilities. Human expertise may improve through practice, while AI performance may change after an update or a shift in the case distribution. An allocation rule that works at the start of a study may therefore lose its advantage. Research should distinguish the effect of a fixed rule from the additional governance needed to monitor and revise it. The current confirmatory experiment tests a specified arrangement over a defined period; adaptive allocation requires its own evaluation design.

Finally, a cumulative program should report comparable standalone benchmarks, intervention costs, and null or adverse findings. Without these elements, apparent differences between studies may reflect incompatible reference conditions rather than genuine boundary conditions. Replications can then ask whether a mechanism travels across settings, rather than merely whether another study produces a favorable average effect. This is the basis for moving from a plausible conceptual model to a more reliable account of when leadership contributes to effective AI governance.

- *Limitations and Future Research*

The framework rests on a focused synthesis and has not been empirically validated. Tasks with verifiable answers fit the proposed experiment more readily than open-ended strategic decisions. Simulation may elicit more attention and expose participants to fewer organizational power constraints than ordinary work. Team contamination, voluntary recruitment of organizations, and system changes during the study will further limit generalizability.

Evaluation criteria may also remain contested. Where no unique answer exists, independent assessors and adjudication are needed, and unresolved disagreement should be reported. Some intervention rights and avenues for recourse depend on actors outside the team. Team leaders cannot necessarily create these arrangements on their own, which limits what their practices can achieve.

Local quality or productivity gains provide no direct evidence of favorable employment or inequality effects. Longitudinal studies would need to examine task redistribution, skill development, compensation, and who captures the gains. Retaining independent judgment is a relevant intermediate outcome, but captures only part of AI's socioeconomic consequences.

VII. CONCLUSION

This article advances a conditional account of leadership as an AI governance capability. Its central argument is that assigning human responsibility becomes consequential only when the decision process gives people a defensible basis for evaluating recommendations and the practical means to act on that evaluation. Extending Ouchen, Chouay, and El Arabi (2026), the model shifts attention from broad leadership configurations to observable practices that may explain differences in oversight. The proposed contribution is therefore an account of how governance is enacted, rather than a claim that leadership necessarily improves AI-supported performance.

The theoretical contribution rests on separating practices, mechanisms, and outcomes. Responsibility boundaries, decision orchestration, and collective learning describe what organizations put in place. Calibrated reliance, actionable contestability, and verified learning feedback explain how those arrangements are expected to affect decisions. Effective oversight, decision quality, and retained unaided competence provide distinct grounds for evaluating their consequences. Maintaining these distinctions avoids defining governance as effective simply because desirable practices are present. It also allows a practice to be implemented faithfully while failing to produce its intended benefit.

The contingency perspective specifies why those benefits may vary. Task evaluability determines whether outcomes can be assessed, while error detectability determines what reviewers can recognize before action. Expertise, available time, evidence access, and incentives shape whether a recognized concern can lead to a justified

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

response. These conditions place limits on the explanatory scope of leadership. Where errors cannot be detected with the available evidence, or intervention rights lie outside the team, additional managerial commitment may be insufficient. Restricting AI use or changing the technical and institutional arrangement may then be more appropriate than adding another approval step.

A further contribution is the separation of human augmentation, human–AI synergy, and procedural protection. Better assisted performance relative to humans alone does not establish superiority to AI alone. Conversely, a configuration that does not achieve synergy may still serve legitimate accountability or recourse functions. Such functions require their own evidence and cannot be inferred from accuracy. The same reasoning applies over time: immediate productivity does not establish that an organization retains the expertise needed to challenge the system, handle exceptions, or recover when support is withdrawn. These distinctions make the evaluation more demanding, but also make its conclusions more useful.

The proposed research program translates the model into comparisons that can support or challenge it. Qualitative fieldwork will refine the interventions and identify barriers to their use. The team-randomized factorial design will estimate the effects of responsibility and orchestration arrangements, while standalone benchmarks will test whether any observed improvement constitutes synergy. A separate randomized learning intervention and delayed unaided assessment will examine retention and transfer. Independent adjudication, preregistered contrasts, and explicit accounting for review resources are intended to prevent favorable perceptions, selective comparisons, or hidden costs from substituting for evidence of effectiveness.

This design nevertheless identifies the effects of specified intervention bundles in particular settings. It does not isolate a universal leadership effect, and randomizing the interventions does not by itself identify their mediating mechanisms. No original empirical results are reported here. Null, adverse, or heterogeneous findings would therefore be substantive tests of the account. Improvements attributable entirely to additional time or technical changes would narrow the leadership explanation; gains in assisted performance accompanied by weaker unaided competence would qualify claims of sustainable benefit. A cumulative research program must preserve these possibilities rather than reinterpret every outcome as confirmation.

For organizations, the model directs attention to the alignment of responsibility with evidence, expertise, authority, and follow-through. The relevant question is whether a concern can be evaluated and resolved, not merely whether a human has signed the decision. The broader scientific task is to establish which governance arrangements improve which outcomes, for whom, at what cost, and over what period. Leadership contributes to effective AI governance to the extent that its practices make these relationships observable, testable, and open to correction. Human–AI complementarity remains a comparative achievement to demonstrate and sustain, rather than an automatic consequence of combining people with a capable system.

- *Scope and Conditions for Empirical Testing*

This article proposes a conceptual model and a protocol for empirical testing. The protocol has not been implemented, no participant data were collected or analyzed, and the proposed intervention effects remain untested. Before recruitment, researchers should establish the applicable ethics review and consent requirements. Deliberately incorrect AI recommendations should be confined to controlled tasks in which they cannot affect real decisions about individuals. Any partial disclosure should be justified in advance and accompanied by an approved debriefing procedure. Where participants are employees, participation decisions and experimental records should be kept separate from performance evaluation and disciplinary decisions. Before collecting confirmatory data, researchers should preregister the primary outcomes, the comparisons needed to test for synergy, and the analysis plan.

REFERENCES

[1]. 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. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (Article 81, pp. 1–16). https://doi.org/10.1145/3411764.3445717

[2]. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at Work. The Quarterly Journal of Economics, 140(2), 889–942. https://doi.org/10.1093/qje/qjae044

[3]. Buçinca, 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. Proceedings of the ACM on Human–Computer Interaction, 5(CSCW1), Article 188, 1–21. https://doi.org/10.1145/3449287

[4]. Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: extension to cluster randomised trials. BMJ, 345, e5661. https://doi.org/10.1136/bmj.e5661

[5]. Edmondson, A. (1999). Psychological Safety and Learning Behavior in Work Teams. Administrative Science Quarterly, 44(2), 350–383. https://doi.org/10.2307/2666999

[6]. Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. Information Systems Research, 33(2), 678–696. https://doi.org/10.1287/isre.2021.1079

[7]. Gonzalez, C., Donahue, K., Goldstein, D. G., Heidari, H., Jalali, M. S., Schelble, B., Singh, A., & Woolley, A. W. (2026). Toward a science of human–AI teaming for decision making: A complementarity framework. PNAS Nexus, 5(3), pgag030. https://doi.org/10.1093/pnasnexus/pgag030

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov456

[8]. 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
[9]. Imai, K., Keele, L., & Tingley, D. (2010). A General Approach to Causal Mediation Analysis. Psychological Methods, 15(4), 309–334. https://doi.org/10.1037/a0020761
[10]. Jaakkola, E. (2020). Designing conceptual articles: four approaches. AMS Review, 10, 18–26. https://doi.org/10.1007/s13162-020-00161-0
[11]. Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. Academy of Management Annals, 14(1), 366–410. https://doi.org/10.5465/annals.2018.0174
[12]. Laux, J. (2024). Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act. AI & Society, 39, 2853–2866. https://doi.org/10.1007/s00146-023-01777-z
[13]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
[14]. National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1
[15]. Orlikowski, W. J. (2000). Using Technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations. Organization Science, 11(4), 404–428. https://doi.org/10.1287/orsc.11.4.404.14600
[16]. Ouchen, I., Chouay, G., & El Arabi, A. (2026). Leadership as a Governance Capability in AI-Enabled Organizations: A Conceptual Framework for Human–AI Complementarity and Socio-Economic Outcomes. International Journal of Innovative Science and Research Technology, 11(3), 723–731. https://doi.org/10.38124/ijisrt/26mar272
[17]. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072
[18]. 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(12), 2293–2303. https://doi.org/10.1038/s41562-024-02024-1