

Emirati Journal of Business, Economics and Social Studies
Vol 5 Issue 2 (2026)
Pages 289 - 304
Available at
www.emiratesscholar.com

Governing AI-Assisted Board Decisions: A Dual-Lock Model for Information Independence and Accountability

Abdulla Saeed Alhebsi
Independent Researcher, United Arab Emirates

ARTICLE INFO

KEYWORDS
artificial intelligence; corporate governance; risk governance; board decisions; delegation; accountability; decision rights; ex-ante commitment

HOW TO CITE
Alhebsi, A. S. (2026). Governing AI-Assisted Board Decisions: A Dual-Lock Model for Information Independence and Accountability. Emirati Journal of Business, Economics and Social Studies, 5(2), 289-304.
https://doi.org/10.54878/ty079w57

ABSTRACT

Purpose: This study develops a governance architecture for board-governed strategic decisions in which artificial intelligence (AI) provides common advice while legal and organizational decision ownership remains human. It examines two linked risks: erosion of independent human information and distortion of delegation through outcome-dependent credit and blame. Methodology/Approach: A theoretical decision-economic model is used to formalize firm-optimal and executive-optimal delegation, an endogenous attribution wedge, common-AI-induced error covariance, and boundary conditions for pre-AI information recording. The analysis treats governance controls as ex-ante commitments rather than ex-post corrections. Findings: Common exposure to an imperfect AI signal can increase correlation among revised human judgments and reduce the effective number of independent checks. Separately, private attribution incentives can move delegation away from the firm-optimal benchmark. The proposed Information Lock records designated human assessments before common AI exposure, while the Accountability Lock fixes effective decision rights and attribution rules before outcomes are known. The model also identifies limits: perfectly reconstructible post-AI judgments reduce the Information Lock mainly to an auditability device, while residual career downside can shift distortion from over-delegation toward rejection or delay of positive-value decisions. Conclusion: The contribution is a unified, selective governance architecture linking information independence and outcome-invariant accountability. It provides boards and senior executives with an implementable framework for preserving informational diversity, clarifying decision ownership, and distinguishing poor process from unfavorable stochastic outcomes in AI-assisted strategic decisions.

Double-blind peer review by independent reviewers.

Received: 20 Jul 2026, Accepted: 27 Oct 2026, Published: 01 Nov 2026

*Corresponding author: abdulla.s.alhebsi@hotmail.com

Copyright: © 2026 by the author. Licensee Emirates Scholar Center for Research & Studies, UAE.
Open access under the Creative Commons Attribution (CC BY) license
https://creativecommons.org/licenses/by/4.0

1. Introduction

The question has become more urgent as artificial intelligence moves from operational analytics into high-level decision processes. In February 2024, Abu Dhabi-based International Holding Company announced Aiden Insight as a non-voting artificial-intelligence observer to its board, positioning the system as a source of data analysis, risk assessment, and strategic insight rather than as a voting decision-maker. In September 2026, the UAE Cabinet launched a Cabinet AI Advisor system built around 32 specialized AI advisers that assess policy and legislative proposals across financial, economic, social, and environmental dimensions and support a governed process for issuing and following up Cabinet decisions. These initiatives do not establish that AI improves board or cabinet performance. They establish a more basic governance fact: a common AI recommendation can now enter consequential decision processes before the human authority that remains accountable has reached a final judgment (Abu Dhabi Media Office, 2024; UAE Government Media Office, 2026).

The economic problem is therefore not whether organizations should simply accept or reject AI. Managerial and Decision Economics has already examined willingness to delegate strategic decisions to algorithms (Schneider & Leyer, 2019), and recent work in the same journal shows that growing trust in AI can reduce strategists' perceived uncertainty while creating the possibility of overreliance on what the system fails to detect (Hutzschenreuter, Kleindienst, & Perkhofer, 2026). A broader behavioral literature documents both algorithm aversion and algorithm appreciation (Dietvorst, Simmons, & Massey, 2015; Logg, Minson, & Moore, 2019; You, Yang, & Li, 2022). The relevant governance question is one of sequencing: what happens to the informational value of multiple human participants when a common AI signal enters before their own assessments are captured?

Advice-timing evidence makes this concern concrete. Yin, Ngiam, Tan, and Teo (2025) find that diagnostic performance improves when AI advice follows an initial human judgment rather than arriving concurrently with the case information, while Balakrishnan, Ferreira, and Tong (2025) show that people can place overly stable weights on algorithmic advice even when they possess private information that should alter optimal reliance. At board level, the concern is not merely individual anchoring. Directors and executives may hold partially non-redundant signals generated by different functions, technologies, geographies, stakeholder relationships, or professional experiences. The economic value of a multi-person decision body depends on that informational diversity, not on headcount alone.

A second leakage concerns decision rights. Delegation is often efficient because it can move a decision toward superior expertise, local information, or capacity (Aghion & Tirole, 1997; Deimen & Szalay, 2019; Ambrus, Baranovskyi, & Kolb, 2021). Yet delegation also changes private attribution. Responsibility can be shifted through delegation and the ability to shift blame can itself motivate delegation (Bartling & Fischbacher, 2012; Hill, 2015). Delegation can facilitate self-interested outcomes (Hamman, Loewenstein, & Weber, 2010), while favorable outcomes create a corresponding problem of who receives credit (Argenton, Potters, & Yang, 2023). These effects become economically important when firm risk and executive career risk are not aligned (Byford, 2017; Gormley & Matsa, 2016; Chari, David, Duru, & Zhao, 2019).

Board-level risk governance is the appropriate locus for this problem even when the board does not personally make every operating decision. Corporate-governance theory distinguishes decision management from decision control (Fama & Jensen, 1983), and boards deliberately allocate reserved matters and delegated authority across the governance hierarchy (Useem & Zelleke, 2006). The G20/OECD Principles likewise connect board responsibility for risk appetite and oversight to clear accountabilities for managing risk and to alignment between strategy and risk management (OECD, 2023). Accordingly, this paper distinguishes board-reserved decisions from board-governed executive decisions: the board may retain final approval for some matters while, for others, it governs the architecture through which executives allocate substantive decision rights.

Existing research provides important pieces of the puzzle but generally treats them separately. Risk-governance scholarship identifies the need to connect strategic risk, governance actors, and adaptive decision making (Stein & Wiedemann, 2016; Stein et al., 2019; Stein et al., 2026). Accountability scholarship distinguishes process from outcome accountability (Siegel-Jacobs & Yates, 1996; Lerner & Tetlock, 1999; de Langhe, van Osselaer, & Wierenga, 2011; Pollmann, Potters, & Trautmann, 2014). Organizational economics explains why decision rights move toward informed agents. Human-AI research shows that advice timing changes information processing. What remains less developed is a single decision-economic model that makes two commitment points explicit: the moment before common AI advice can reshape human judgments, and the moment before an outcome can reshape the attribution of decision ownership.

This paper develops a Dual-Lock Risk Governance Model around those two commitment points. The Information Lock time-stamps designated human assessments before common AI exposure while allowing later revision. The Accountability Lock fixes effective decision rights, delegation rationale, outcome attribution, and process accountability before the stochastic outcome is known. The second lock does not make a delegator responsible for every bad outcome and does not punish legitimate delegation. Instead, it removes excess outcome attribution that is inconsistent with the pre-specified rights structure while preserving accountability for the quality of the delegation process.

The paper makes three contributions. First, it endogenizes the delegation benchmark by linking the firm's optimal delegation share to the decision's success probability, upside, downside, capacity benefit, and coordination cost. Private success-credit retention and excess blame shifting then generate an attribution wedge at that benchmark rather than entering as an assumed linear rent. The analysis also identifies a second-order agency problem: once excess blame shifting is removed, residual outcome-contingent career downside can redirect distortion from over-delegation toward rejection or delay of a positive-value decision. Second, the paper derives a common-AI correlation result and clarifies the informational boundary of pre-AI recording. When multiple humans blend their private judgments with the same imperfect AI signal, their reported errors become more correlated and the effective number of independent human checks falls; when the post-exposure transformation is perfectly invertible, however, a pre-AI record adds auditability rather than new statistical information. An information-acquisition extension shows why the lock can still have substantive informational value when governance requires separately identifiable private signal production. Third, the paper reframes the two controls as a unified ex-ante commitment architecture within risk governance. One lock preserves the baseline information set before common influence; the other preserves the rights-and-accountability map before realized outcomes can reshape organizational narratives.

The central research question is therefore: How do ex-ante information and accountability commitments affect the informational diversity and incentive alignment of AI-assisted, board-governed strategic decisions? The remainder of the paper develops the risk-governance and decision-rights foundations, formalizes the delegation and information mechanisms, derives the core results and boundary conditions, and translates them into selective governance implications for boards and senior decision authorities.

Purpose and Objectives

The study has three objectives: (1) to formalize how common AI advice can affect the independence of multiple human judgments; (2) to model how private credit, blame, and career incentives can distort delegation away from the firm-optimal allocation of decision rights; and (3) to evaluate a unified ex-ante governance architecture that addresses both mechanisms while identifying the conditions under which each control adds value. The central research question is: How do ex-ante information and accountability commitments affect informational diversity and incentive alignment in AI-assisted, board-governed strategic decisions?

2. Literature Review and Theoretical Foundations

2.1 Risk Governance as Strategic Decision Architecture

Risk governance is not synonymous with risk management. Stein and Wiedemann (2016) distinguish the concept precisely because specialized risk-management instruments and conventional corporate-governance structures can remain fragmented in dynamic environments. Their formulation positions risk governance as an overarching bridge that organizes collaborative processes of interaction and decision making among actors involved in the collective problem of strategic risk, with long-term value optimization as the governing objective. This distinction matters for the present paper because the unit of analysis is not a risk register or control function; it is the architecture of a material decision through which risk, opportunity, information, and authority are combined.

2.2 Boards, Reserved Matters, and Delegable Decision Rights

Board-level risk governance does not imply that a board should personally make every risky decision. Fama and Jensen (1983) distinguish decision management from decision control, while Useem and Zelleke (2006) show that boards deliberately decide what to retain and what to delegate. We therefore distinguish two classes of consequential decisions. In a board-reserved matter, the governing body retains a non-delegable approval or veto right, although analysis, recommendation, negotiation, and execution rights may be allocated elsewhere. In a board-governed executive decision, the board establishes the risk appetite, delegation framework, and

ISSN: 2791-3171 - https://doi.org/10.54878/ty079w57

accountability architecture while a senior executive is the effective substantive decision owner.

2.3 Delegation as Expertise and Attribution

Delegation can create firm value by moving rights toward superior expertise, local information, or scarce capacity. Formal models of delegated expertise show that authority and communication should respond to information acquisition and preference alignment rather than hierarchy alone (Aghion & Tirole, 1997; Deimen & Szalay, 2019; Ambrus et al., 2021). The model therefore begins from a firm-optimal delegation benchmark rather than from a presumption that centralization is desirable.

Attribution creates a second motive. Experimental evidence shows that responsibility can be shifted through delegation and that blame avoidance can itself motivate the transfer of control (Bartling & Fischbacher, 2012; Hill, 2015). Delegation can facilitate self-interested choices (Hamman et al., 2010), while favorable outcomes generate a parallel credit-allocation problem (Argenton et al., 2023). The resulting agency problem is not delegation per se but excess attribution: private success credit retained beyond transferred rights and failure accountability shifted beyond what the same rights transfer justifies.

2.4 Outcome-Invariant Accountability as Ex-Ante Commitment

A governance response cannot simply make the delegator substantively responsible for every eventual outcome. Such a rule would erase the value of genuine authority transfer and could discourage delegation to better-informed specialists. The proposed accountability lock instead separates outcome attribution from delegation-process accountability. Outcome attribution follows the effective rights fixed ex ante. Process accountability concerns whether the transfer itself had an adequate competence, capacity, urgency, or information rationale and whether its scope was documented. The process test is applied under the same rule whether the project later succeeds or fails.

2.5 AI Advice as a Common Signal

Artificial intelligence creates a different commitment problem. An AI adviser can alter the information environment without possessing any formal decision right. Yin et al. (2025) show that the timing of AI advice changes diagnostic decision making, and Balakrishnan et al. (2025) show that human decision-makers can apply overly stable weights to algorithmic advice despite private information. At board level, a common recommendation is especially important because the organizational value of multiple participants comes partly from non-redundant information.

2.6 Theoretical Gap and Dual Ex-Ante Commitment Logic

Risk-governance research calls for coordinated, strategic, cross-functional control of risk; organizational economics explains the movement of decision rights; accountability research explains process and outcome responsibility; and human-AI research explains advice-taking. Yet two adjacent moments in a consequential decision remain weakly integrated: before a common adviser can reshape the information set and before a realized outcome can reshape the attribution map. Both moments create an option for retrospective reinterpretation. The first can obscure what participants believed independently; the second can obscure who actually owned the decision.

The Dual-Lock model treats these as two applications of the same governance principle: commit the baseline record before the event that can make retrospective rewriting privately or cognitively attractive. The locks need not be behaviorally synergistic. Their integration is architectural. The board or senior authority establishes a workflow that preserves pre-exposure information and pre-outcome decision rights, then allows updating, deliberation, and legitimate delegation inside those fixed reference points. This is the transition from risk governance as a broad philosophy to a testable decision-economic mechanism.

3. Methodology and Formal Research Approach

3.1 Actors, Decision Classes, and Timing

Consider a material strategic decision governed by a board-level risk architecture. There are four roles: a governing board G, a senior human authority E, a potential human delegate D, and an AI adviser A. The board determines whether the matter is board-reserved or management-delegable and establishes the applicable risk appetite, delegation boundary, and accountability process. The AI adviser supplies information but has no formal decision right in the model.

Let λ in $[0, \bar{\lambda}]$ denote the share of the delegable substantive decision-right bundle transferred from E to D. The upper bound $\bar{\lambda}$ is itself a governance parameter. For a management-delegable matter, $\bar{\lambda}$ may equal 1. For a board-reserved matter, $\bar{\lambda}$ is below 1 because the board retains a non-delegable approval or veto right even if recommendation, negotiation, analytical, or execution rights are transferred. Thus λ measures effective delegated authority within the legally and organizationally permissible set rather than formal title alone.

The timeline is ex ante. Stage 1 fixes the information architecture and records any designated independent human assessments. Stage 2 allocates effective decision rights. Stage 3 permits AI exposure, deliberation, revision, and the final human decision. Stage 4 realizes the stochastic outcome. The Information Lock operates before common AI exposure in Stage 3; the Accountability Lock operates before Stage 4. Both are commitment devices, but they protect different objects.

3.2 Firm-Optimal Delegation with Endogenous Decision Stakes

Let $p(\lambda) = p_0 + h\lambda$ be the probability of a favorable outcome, where p_0 is the baseline success probability and h is the marginal effect of moving delegable rights toward D. A positive h captures a competence or local-information advantage of the delegate; a negative h captures a competence disadvantage. Parameter values are restricted so that $0 < p(\lambda) < 1$ on the feasible interval. Let $G > 0$ be the firm's gain from success, $L > 0$ its loss from failure, s be a linear capacity or speed benefit from delegation, c be a linear transfer cost, and $k > 0$ the convex coordination and control cost.

$$F(\lambda) = p(\lambda)G - [1 - p(\lambda)]L + (s - c)\lambda - (k/2)\lambda^2. \quad (1)$$

Relative to retaining the decision right, the marginal firm benefit of delegation is

$$BR = h(G + L) + s - c. \quad (2)$$

The notation makes the risk economics explicit. Delegation is more valuable when a superior delegate increases the probability of success on a decision with a large combined upside and downside, or when capacity benefits are high. It is less valuable when transfer and coordination costs dominate.

Lemma 1 - Unique firm-optimal delegation. For $k > 0$, $F(\lambda)$ is strictly concave. The firm-optimal delegation share is

$$\lambda F^* = \min\{\bar{\lambda}, \max\{0, BR/k\}\}. \quad (3)$$

Proof. $F'(\lambda) = BR - k\lambda$ and $F''(\lambda) = -k < 0$. Strict concavity gives a unique unconstrained maximizer BR/k, which is projected onto the feasible interval $[0, \bar{\lambda}]$.

Lemma 1 embeds the board boundary directly. A high value of λ is not evidence of governance failure when expertise, speed, or local information makes delegation firm-optimal. A board-reserved matter simply truncates the feasible transfer at $\bar{\lambda}$ rather than forcing the model to treat the board as an operational decision-maker.

3.3 Executive-Optimal Delegation and the Endogenous Attribution Wedge

Let α in (0,1] be the fraction of firm value internalized by E through aligned compensation, reputation, ownership, or career concerns tied to organizational performance. In addition, let $BC > 0$ denote the magnitude of private success credit and $AC > 0$ the magnitude of private failure-accountability cost that can be redistributed beyond what the rights structure justifies. Let $r \geq 0$ measure excess success credit retained by E per unit of delegated authority and $b \geq 0$ measure excess failure accountability shifted away from E per unit of delegated authority. These are excess-attribution parameters: legitimate changes in credit and accountability that follow the pre-specified rights structure are not counted as rents.

The private attribution payoff from delegation is

$$\Phi(\lambda) = p(\lambda)\, BC\, r\, \lambda + [1 - p(\lambda)]\, AC\, b\, \lambda. \quad (4)$$

and the executive maximizes

$$U(\lambda) = \alpha F(\lambda) + \Phi(\lambda). \quad (5)$$

The marginal private attribution incentive is

$$\Gamma(\lambda) = \Phi'(\lambda) = BC\, r[p_0 + 2h\lambda] + AC\, b[1 - p_0 - 2h\lambda]. \quad (6)$$

Assume U is strictly concave on the relevant interior region, which requires $KE = \alpha k - 2h(BC\, r - AC\, b) > 0$. The executive's unconstrained optimum is

$$\lambda E^* = [\alpha BR + BC\, r\, p_0 + AC\, b(1 - p_0)] / KE. \quad (7)$$

Proposition 1 - Endogenous attribution wedge. Suppose λF^* and λE^* are interior and $KE > 0$. Then

$$\lambda E^* - \lambda F^* = \Gamma(\lambda F^*) / KE. \quad (8)$$

Accordingly, executive-optimal delegation exceeds the firm optimum if and only if the marginal private attribution payoff evaluated at the firm optimum is positive.

Proof. Because $F'(\lambda F^*) = 0$, $U'(\lambda F^*) = \Gamma(\lambda F^*)$. Under strict concavity, $U'(\lambda) = KE(\lambda E^* - \lambda)$ in the interior. Evaluating at λF^* gives Equation (8).

The result does not build self-protective delegation into a reduced-form linear rent. The wedge is generated by the decision's success probability, the way delegation changes that probability, the magnitude of personal upside and downside, and the organization's attribution rules. Career risk matters most when failure remains sufficiently likely and excess blame shifting b is valuable; credit capture matters when success remains likely and r is positive.

3.4 Why Removing Delegator Credit Alone Is Not Sufficient

A strict full-delegation rule may assign substantive outcome credit entirely to D. In the model, this removes excess retained success credit by setting $r = 0$. It does not,

by itself, eliminate the option to shift excess failure accountability if b remains positive.

Corollary 1 - Zero-credit insufficiency. With r = 0, the marginal attribution wedge at the firm optimum becomes

$$\Gamma 0(\lambda F^*) = AC\, b[1 - p_0 - 2h\lambda F^*]. \quad (9)$$

Whenever the bracketed term is positive and b > 0, the executive still prefers more delegation at λF^* despite receiving no excess success credit. A zero-credit rule therefore addresses only one side of the attribution asymmetry.

The governance implication is the same as the practical observation motivating the model: outcome credit should follow the effective decision owner, but the original authority also needs an ex-ante, outcome-invariant responsibility for the quality and legitimacy of the transfer itself. Otherwise a leader can lose the upside-credit option yet retain the downside-insurance option.

3.5 Accountability Lock and Baseline Alignment Under Restricted Private Payoffs

The Accountability Lock fixes the effective rights structure before the outcome and prohibits excess attribution inconsistent with that structure. For compliant delegation, the lock sets r = 0 and b = 0. It does not impose a generic penalty on delegation. Process accountability is triggered only by a documented defect in the transfer—such as an unauthorized scope, an unqualified delegate, or a missing required consultation—and is evaluated under the same rule regardless of the realized outcome.

Proposition 2 - Baseline alignment absent residual private career exposure. If r = b = 0 for a compliant delegation, no process penalty is imposed merely because delegation occurred, and E has no outcome-contingent private payoff beyond the aligned component $\alpha F(\lambda)$, then $U(\lambda) = \alpha F(\lambda)$. Therefore the executive and firm delegation optima coincide on the feasible interval:

$$\lambda L^* = \lambda F^*. \quad (10)$$

Proof. With r = b = 0, $\Phi(\lambda) = 0$ for all λ. Since $\alpha > 0$, maximizing $U(\lambda) = \alpha F(\lambda)$ is equivalent to maximizing $F(\lambda)$.

This result clarifies the intended governance design, but it is a baseline rather than a universal alignment theorem. The lock is not 'more accountability' in the abstract. It makes outcome attribution consistent with pre-specified rights while evaluating delegation quality separately. Legitimate expert delegation is not taxed; excess retrospective credit and blame reallocation are removed. Section 3.6 relaxes the additional assumption that no residual outcome-contingent career penalty remains.

3.6 Residual Career Risk and the Rejection-Substitution Boundary

Removing excess attribution does not guarantee that a senior authority will implement every positive-value decision. A leader may still face a legitimate or institutionally imposed personal downside from authorizing a risky decision that later fails. Let $R \geq 0$ denote the outcome-contingent career or reputational loss borne by E when the organization undertakes the decision and the unfavorable state occurs, after excess blame shifting has already been removed. This residual exposure is distinct from b: b captures unjustified reallocation of accountability through delegation, whereas R captures the remaining private consequence of owning or approving a risky decision.

Evaluate the adoption decision at the firm-optimal feasible delegation share λF^*. If the outside option from rejecting or indefinitely deferring the proposal is normalized to zero, the executive's expected utility from undertaking the decision under the Accountability Lock is

$$UTake = \alpha F(\lambda F^*) - [1 - p(\lambda F^*)]R. \quad (11)$$

A project can therefore have positive firm value, $F(\lambda F^*) > 0$, and still be privately unattractive to the decision owner when residual career downside is sufficiently large.

Corollary 2 - Rejection substitution. Suppose $F(\lambda F^*) > 0$ and $R > 0$. After excess credit and blame reallocation are removed, E rejects or indefinitely defers the positive-value decision whenever

$$R > \alpha F(\lambda F^*) / [1 - p(\lambda F^*)]. \quad (12)$$

The implication is important: an Accountability Lock can eliminate self-protective over-delegation yet reveal or redirect a separate career-risk distortion into excessive caution. A governance system that merely makes blame harder to shift can therefore produce paralysis if it does not also distinguish bad process from bad luck.

Let $\chi \in [0,1]$ denote an ex-ante process-protection parameter: the fraction of outcome-contingent career penalty that the board or governing authority commits not to impose when the decision was within approved risk appetite, the required process was followed, material dissent was disclosed, and no misconduct or concealed information is later identified. Residual downside becomes $(1 - \chi)R$. The minimum protection required to make the positive-value decision privately acceptable is

$$\chi min = \max\{0,\ 1 - \alpha F(\lambda F^*) / ([1 - p(\lambda F^*)]R)\}. \quad (13)$$

This is not immunity from accountability. It is an outcome-invariant commitment to evaluate a compliant risky decision on the pre-specified process and risk mandate rather than treating an unfavorable realization

as proof that the decision was improper. The distinction is consistent with evidence that process and outcome accountability can affect judgment differently (de Langhe, van Osselaer, & Wierenga, 2011; Pollmann, Potters, & Trautmann, 2014) and with evidence that managerial career incentives can generate value-destroying risk avoidance (Gormley & Matsa, 2016).

3.7 Human Judgments Before and After a Common AI Signal

Now consider the Information Lock. Let the unknown decision-relevant state be θ. Before AI exposure, human participant i holds an unbiased private assessment

$$x_i = \theta + \varepsilon_i, \mathrm{Var}(\varepsilon_i) = \sigma_H^2, \mathrm{Corr}(\varepsilon_i, \varepsilon_j) = \rho_0 < 1. \quad (14)$$

The common AI adviser produces

$$y = \theta + \eta, \mathrm{Var}(\eta) = \sigma_A^2. \quad (15)$$

For the baseline derivation, η is independent of the pre-exposure human errors. After seeing the same AI recommendation, each participant reports a revised continuous judgment

$$z_i = (1 - w)x_i + wy, \; 0 < w < 1. \quad (16)$$

where w is the common weight placed on the AI signal. The assumption of a common w is a transparent benchmark; heterogeneous weights are a natural extension. The error variance of each revised human report is

$$\sigma_Z^2 = (1 - w)^2\sigma_H^2 + w^2\sigma_A^2, \quad (17)$$

while the covariance between two revised human reports is

$$\mathrm{Cov}(z_i - \theta, z_j - \theta) = (1 - w)^2\rho_0\sigma_H^2 + w^2\sigma_A^2. \quad (18)$$

Proposition 3 - Common-AI induced correlation. Under Equations (14)-(18), the pairwise error correlation among revised human judgments is

$$\rho_1 = [(1 - w)^2\rho_0\sigma_H^2 + w^2\sigma_A^2] / [(1 - w)^2\sigma_H^2 + w^2\sigma_A^2], \quad (19)$$

and

$$\rho_1 - \rho_0 = [w^2\sigma_A^2(1 - \rho_0)] / \sigma_Z^2 > 0. \quad (20)$$

Proof. Substitute Equations (17) and (18) into the correlation definition and subtract ρ_0. The remaining numerator is $w^2\sigma_A^2(1 - \rho_0)$, which is strictly positive for an imperfect common AI signal, positive AI weight, and non-identical pre-exposure human errors.

The result is deliberately not a claim that AI advice worsens individual accuracy. A high-quality AI can reduce σ_Z^2. The proposition instead identifies a governance externality of common advice: revised human reports share the AI's error component, so agreement among them is less independent than the same headcount suggests.

3.8 Effective Number of Independent Human Checks

A convenient way to express the board-level implication is through the effective number of independent human checks. For N equal-variance reports with common pairwise correlation ρ, define

$$N_{eff}(\rho) = N / [1 + (N - 1)\rho]. \quad (21)$$

This is the number of independent equal-variance observations that would generate the same variance of the sample mean. Proposition 4 - Independence loss under common AI exposure. Under Proposition 3,

$$N_{eff}(\rho_1) < N_{eff}(\rho_0). \quad (22)$$

Proof. Neff is strictly decreasing in ρ for $N > 1$, and Proposition 3 establishes $\rho_1 > \rho_0$.

This is the precise informational rationale for the lock. Capturing x_i before AI exposure preserves an auditable set of pre-exposure judgments with effective independence $N_{eff}(\rho_0)$. The organization may still use the revised judgments z_i and the AI signal y. What it should not do is treat N correlated post-exposure endorsements as if they were N independent confirmations. The lock preserves the baseline against which influence and convergence can be measured.

3.9 Reconstruction Boundary and Independent Information Production

A natural boundary question arises from Equation (16): if the organization observes the post-AI report z_i, the AI signal y, and the exact weight w, is a separate record of x_i informationally necessary? Under the linear truthful-reporting benchmark, the answer is no.

Boundary Result 1 - Perfect reconstruction. If $0 < w < 1$ is known, y is observed, reporting is truthful and continuous, and Equation (16) holds exactly, then the pre-exposure assessment can be recovered as

$$x_i = (z_i - wy) / (1 - w). \quad (23)$$

In that restrictive benchmark, the Information Lock does not create additional statistical information. Its value is auditability: it prevents N post-exposure endorsements from being represented as N independent confirmations and provides an observed baseline for measuring influence. The stronger claim that pre-AI capture preserves information requires a friction that makes reconstruction incomplete or changes what private information is produced in the first place.

Such frictions are plausible in board settings: AI weights are rarely observed as stable scalar parameters; final recommendations can be categorical rather than

continuous; deliberation can change remembered reasons; participants may strategically conform; and exposure to a common recommendation may alter how much effort is devoted to developing a private assessment. The following extension formalizes the last mechanism without claiming that it is universal.

Let each participant choose private information-acquisition effort $e \geq 0$. The variance of the private signal is $\sigma_H^2(e) = 1/(\tau_0 + qe)$, with $\tau_0 > 0$ and $q > 0$, and effort cost is $(\kappa/2)e^2$ with $\kappa > 0$. Under a locked architecture, the separately identified private assessment remains a first-class input, so the participant minimizes $JL(e) = \sigma_H^2(e) + (\kappa/2)e^2$. If only the post-AI blended report is organizationally used, the human-specific error enters that report with weight $(1 - w)$, so effort relevant to the human component solves $JU(e) = (1 - w)^2\sigma_H^2(e) + (\kappa/2)e^2$, ignoring terms independent of e.

Proposition 5 - Common-advice crowd-out of private information acquisition. Let eL and eU be the unique interior effort choices under JL and JU. Then for $0 < w < 1$,

$eU < eL$. (24)

Proof. The first-order conditions can be written as $\kappa eL(\tau_0 + qeL)^2 = q$ and $\kappa eU(\tau_0 + qeU)^2 = (1 - w)^2 q$. The function $f(e) = \kappa e(\tau_0 + qe)^2$ is strictly increasing for $e \geq 0$. Because $(1 - w)^2 q < q$, the second equation has the smaller solution.

Proposition 5 gives the Information Lock a substantive role beyond archival reconstruction under one explicit mechanism: when governance preserves and uses a separately identifiable human signal, it can sustain incentives to produce private information that would otherwise be discounted inside a common blended report. The result should be read as an economic extension, not as a universal behavioral claim about every director or every AI system.

3.10 Two Locks as a Unified Ex-Ante Commitment Architecture

The delegation and information results are not assumed to interact multiplicatively. Their theoretical unity is temporal and institutional. In each case the organization faces an event that can make a valuable baseline difficult to reconstruct or easy to reinterpret: common advice can reshape both reported judgment and incentives to produce separate private information, while realized outcomes can reshape attribution and career evaluation. The governance response is to preserve the relevant baseline before that event. The Information Lock fixes the pre-influence assessment record; the Accountability Lock fixes the pre-outcome rights-and-attribution record and, where appropriate, the process-based protection attached to an approved risky decision.

This common commitment logic explains why the controls belong inside risk governance rather than as unrelated technical procedures. Risk governance already aims to connect strategic risk, actors, information, and accountability across organizational levels. The Dual-Lock model specifies concrete reference points that make that coordination auditable at the decision level. The boundary results also prevent overclaiming: a perfectly invertible post-AI record reduces the incremental informational value of pre-AI capture, and an attribution lock without process-based protection can redirect agency conflict from over-delegation toward excessive rejection when legitimate career downside is large. Selective activation therefore depends on the particular information and incentive frictions present in the decision.

3.11 Boundary Conditions

If post-AI reports are truthful, continuous, and perfectly invertible because y and w are known, the Information Lock adds auditability but no new statistical information; its stronger information-preservation role depends on non-invertibility, strategic reporting, path dependence, or endogenous private-signal production.

If residual outcome-contingent career downside R remains large after excess attribution is removed, the Accountability Lock can shift distortion from over-delegation to rejection or delay. Process-based risk acceptance is then required to prevent bad luck from being treated as evidence of bad process.

If $h(G + L) + s - c \leq 0$, the firm benchmark is no delegation unless the feasible set or omitted benefits change. Positive delegation in this region is therefore especially informative about non-firm motives.

If $h(G + L) + s - c \geq k\lambda$, the firm prefers the maximum feasible delegation. The architecture should not force symbolic centralization.

If $r = b = 0$ before the Accountability Lock is applied, the baseline model contains no attribution wedge and the lock adds traceability rather than incentive alignment.

If AI is nearly error-free, σ_A^2 approaches zero and the induced error-correlation term in Equation (17) vanishes even though participants may converge on the correct signal.

If human pre-exposure judgments are already perfectly correlated (ρ_0 approaches 1), preserving additional headcount adds little independent information.

If AI advice weights differ materially across participants, the common-weight model becomes a benchmark rather than an exact description; the same common-error

mechanism remains but pairwise covariances become heterogeneous.

If delay cost is extreme, such as an emergency shutdown or rapidly propagating cyber event, the full sequence may be dominated by a parallel or abbreviated control mode.

3.12 Formal Results Summary

Table 1. summarizes the formal results and the principal boundary results. The materiality rule used later is presented as an implementation decision rule rather than as a core theoretical contribution.

Result	Economic mechanism	Governance implication
Lemma 1	Firm-optimal delegation is driven by decision stakes, the delegate effect on success probability, capacity/speed benefits, transfer cost, and convex coordination cost.	Delegation can rationally be zero, partial, or maximal within the board-defined feasible rights boundary.
Proposition 1	Excess success-credit retention and excess blame shifting generate a marginal attribution payoff at the firm optimum.	Self-protective delegation is defined relative to the firm benchmark, not by high delegation itself.
Corollary 1	Removing excess delegator success credit does not eliminate the wedge when excess blame shifting remains valuable.	Credit rules alone are insufficient; attribution and process accountability must be fixed ex ante.
Proposition 2	When compliant delegation has no excess credit/blame reallocation, executive and firm delegation optima coincide in the baseline model.	Outcome attribution should follow effective rights, while delegation quality is evaluated separately and outcome-invariantly.
Propositions 3-4	Blending private human judgments with the same imperfect AI signal induces common error correlation and lowers the effective number of independent human checks.	Capture valuable human assessments before common AI exposure; do not treat post-exposure headcount as independent confirmation.
Implementation Rule	Use the relevant lock only when expected recoverable information or incentive value exceeds administrative and delay costs.	Apply selective modes rather than universal escalation.
Corollary 2	Residual outcome-contingent career downside can redirect distortion from excessive delegation to rejection or delay of a positive-value decision.	Outcome-invariant attribution should be paired with ex-ante process-based risk acceptance when career downside would otherwise produce paralysis.
Boundary Result 1 / Proposition 5	Perfectly invertible post-AI reports make pre-AI capture informationally redundant beyond auditability; when only blended reports are used, common advice can reduce incentives to acquire separately identifiable private information.	Use the Information Lock where reconstruction is incomplete or independent information production is itself valuable; do not mandate it mechanically.

4. Findings and Governance Architecture

The economic model implies a governance architecture that differs from both "AI decides" and "AI merely summarizes." The board or senior authority remains the legal and organizational decision owner, but the process is designed so that useful independent information is not destroyed before it can be observed and so that the distribution of decision rights cannot be rewritten after the outcome becomes known. The recommendations below are deliberately subordinate to the theory: each implements a mechanism derived above rather than adding unrelated best practices.

4.1 Where Independent Signals Matter, Capture Them Before Common AI Advice

The Information Lock is not a prohibition on AI-first analysis by the system itself. It is a sequencing rule for designated human contributions when their independent information has material value. In the restrictive case where the exact pre-AI assessment can be reconstructed from complete system logs, the lock's incremental role is auditability and correct interpretation of correlated endorsements. Its stronger informational value arises when AI exposure changes private-information production, reporting, memory, or strategic conformity. Organizations should therefore activate the lock only where those frictions are plausible and the independent signal is worth preserving.

After AI exposure, each participant may revise the judgment. The revised view is often more decision-relevant than the initial one; however, the pair of observations is more informative than the revision alone because it reveals where the AI changed a professional judgment and why. This sequencing is consistent with experimental evidence on AI-advice timing (Yin et al., 2025) while adapting the logic to a multi-person governance environment.

4.2 AI Should Analyze Dissent, Not Erase It

The second AI pass should not simply convert the board into a majority vote or generate a polished consensus paragraph. Its role is to identify unresolved assumptions, non-overlapping information, changes between pre- and post-AI judgments, and scenarios under which minority views become decisive. A dissenting specialist may be wrong, but the organization should know whether the disagreement is caused by a different risk appetite, different factual information, or a different causal model.

The governance benefit is greatest when AI is used as a structured comparison engine rather than as an authority signal. The final decision-maker can then see not only what the model recommends, but how human experts changed after exposure to it. This makes AI influence auditable without treating every change of mind as bias.

4.3 Decision Rights Must Be Recorded in Effective Terms

Delegation documents frequently state that authority is transferred while preserving broad approval, veto, budget, or override rights at a higher level. The model therefore uses effective λ and the board-defined upper bound $\bar{\lambda}$ rather than the binary label 'delegated.' Governance records should identify which rights are legally or organizationally non-delegable, who can finally approve, who can veto or reverse, whose signature is required, and who controls the resources needed to execute the choice. Outcome attribution should follow this effective rights structure rather than hierarchy or rhetorical ownership.

4.4 Full Delegation Transfers Substantive Outcome Credit to the Delegate

For full delegation, the strict zero-outcome-credit rule is retained: the original authority holder should not claim the outcome credit merely because the authority was originally theirs to transfer. The senior leader can legitimately receive separate recognition for building a capable team or designing an effective delegation system, but that is not the same object as credit for the delegated decision. Keeping those categories separate removes the "heads I win, tails you lose" attribution option.

The symmetry matters. If the delegate owns the decision within the recorded scope, the delegate also bears decision accountability within that scope. The delegator retains accountability for the quality and legitimacy of the delegation act. This two-level structure protects genuine empowerment while denying the superior a post-outcome choice among competing ownership narratives.

4.5 Evaluate Delegation Process Ex Ante, Not the Luck of the Outcome

A successful project does not prove that delegation was prudent, and a failed project does not prove that it was negligent. Risky strategic decisions contain irreducible outcome variance. Process accountability should therefore ask questions that can be answered before the outcome: Was the delegate competent for this class of decision? Was the scope clear? Were relevant information and resources transferred? Were retained approval rights explicitly stated? Was the delegation motivated by a defensible capacity, expertise, or urgency rationale?

This distinction responds to the accountability literature, including evidence that decision-based and outcome-based accountability can affect risk-taking differently (Pollmann, Potters, & Trautmann, 2014). It also prevents governance from becoming hindsight punishment, which would encourage risk avoidance rather than value-maximizing risk governance.

The same rule must also protect the organization from excessive caution. If a material decision was within approved risk appetite, the required analysis and consultation occurred, dissent was disclosed, and no misconduct is identified, an unfavorable realization should not by itself trigger a retrospective career penalty. Recording that process-based protection ex ante reduces the incentive to replace self-protective delegation with self-protective rejection.

4.6 Preserve the Decision Record After the Outcome

The record should be immutable in the narrow governance sense: later commentary can be added, but the original decision-right map, pre-AI judgments, AI recommendation, revised judgments, delegation scope, and final approval should not be overwritten. This protects organizational learning. It also enables later review of whether AI changed views beneficially, whether dissent contained useful private information, and whether delegation was justified at the time it occurred.

4.7 Ex-Ante Decision Rights Record

A practical implementation of the accountability lock requires a concise Ex-Ante Decision Rights Record. The record is not a claim that any regulator mandates these exact fields; it is a research artifact derived from the model. It separates the right to decide from responsibility for designing a delegation process, identifies any board-reserved right, and preserves the pre-outcome allocation against retrospective reinterpretation. The minimum record is summarized in Table 2.

Table 2
Ex-Ante Decision Rights Record: minimum pre-outcome record

Field	Minimum record
Decision and scope	Material decision, alternatives, scope, and date.
Original authority holder	Person or body that formally possessed the decision right before any delegation.
Effective decision owner	Person or body with the substantive final decision right after retained veto, concurrence, approval, and override rights are considered.
Delegable rights / retained rights	Effective transfer within the permitted bundle, including any board-reserved approval, veto, signature, budget, concurrence, or override right.
Delegation rationale	Ex-ante competence, capacity, urgency, specialization, or information rationale.
Pre-AI human assessment	Initial recommendation, core assumptions, material risks, and reversal conditions.
AI assessment	Recommendation, assumptions, uncertainty, and material information dependencies.
Post-AI revision	Whether the human assessment changed after AI exposure and why.
Outcome credit rule	Who receives substantive decision credit under the pre-specified rights structure; full transfer of a substantive decision right does not permit excess delegator credit.
Process accountability	Who remains answerable for the quality, scope, delegate competence, and procedural legitimacy of the delegation decision, evaluated independently of outcome luck.
Audit preservation	Original entries are not overwritten after the outcome; later commentary is appended separately.
Risk-acceptance / process protection	Whether a procedurally compliant decision within approved risk appetite is protected from adverse retrospective judgment based solely on outcome; list exceptions for misconduct, concealment, unauthorized scope, or process breach.

The model does not support universal escalation. Let ΔVI denote the expected value of preserving an independent pre-AI baseline, ΔVA the expected value of removing a material attribution distortion, τ the direct administrative cost of the workflow, and ωt the economic cost of incremental delay. A board or senior authority should activate the relevant lock or locks only when the expected recoverable value exceeds those costs, as summarized in Equation (25).

$$NB = \Delta VI + \Delta VA - \tau - \omega t. \quad (25)$$

This is an implementation rule, not a separate theoretical theorem. It translates the model into selective governance modes. Routine and reversible decisions may require neither lock; information-sensitive decisions may require only the Information Lock; delegation-sensitive decisions may require only the Accountability Lock; and high-stake decisions containing both problems may justify the full Dual-Lock workflow. Table 3 summarizes the modes.

Table 3
Selective governance modes under the Dual-Lock architecture

Mode	Information Lock	Accountability Lock	Typical use
0 – Standard	No	No	Routine, reversible, low-stake decisions with little attribution ambiguity.
1 – Information	Yes	No	High informational diversity or material common-AI influence, but little delegation distortion.
2 – Accountability	No	Yes	Delegation-sensitive decision with limited information-independence value.
3 – Dual-Lock	Yes	Yes	High-stake decisions with valuable distributed information and material attribution incentives under delegated authority.

5. Discussion and Managerial Implications

The discussion translates the formal results into a deliberately bounded contribution. The paper does not claim that delegation, accountability, AI advice, or board-level AI are individually new phenomena. Its contribution lies in specifying two ex-ante commitment failures inside board-governed risk decisions: a common

adviser can reduce the independence represented by subsequent human agreement, and outcome-contingent attribution can distort the allocation of delegable rights. The model derives each mechanism separately and integrates them through a common risk-governance architecture without assuming a behavioral synergy that has not been demonstrated.

5.1 Risk Governance: From Risk Registers to Decision Architecture

The first contribution is to risk governance. Prior work positions risk governance as a bridge between corporate governance and risk management and emphasizes proactive adaptation and coordinated executive accountability (Stein & Wiedemann, 2016; Stein et al., 2019; Stein et al., 2026). The Dual-Lock model moves this logic to two commitment points inside a consequential decision: preserving designated independent information before common influence and preserving the rights-and-attribution map before outcome realization. A sophisticated risk system can still lose value if dissent is formed only after a common recommendation or if credit and blame are reconstructed after results are known. The model therefore treats decision rights, information flows, and pre-committed accountability as core elements of risk governance.

5.2 Delegation Theory: Legitimate Delegation Versus Attribution-Induced Delegation

The model does not treat extensive delegation as evidence of weak leadership. Firm-optimal delegation depends on risk stakes, delegate competence, capacity benefits, transfer and coordination costs, and the feasible rights boundary. Distortion arises only when private attribution moves the executive away from that benchmark. Proposition 1 identifies this executive-firm gap through the marginal private value of excess success-credit retention and blame shifting, connecting efficient expertise transfer with evidence that delegation can alter responsibility and private incentives (Aghion & Tirole, 1997; Bartling & Fischbacher, 2012; Hamman et al., 2010; Hill, 2015; Deimen & Szalay, 2019; Ambrus et al., 2021; Argenton et al., 2023). Because identical delegation shares can reflect different motives, an ex-ante record of rights and rationale is analytically useful, not merely administrative.

5.3 Accountability: Outcome-Invariant Rather Than Outcome-Blind

The Accountability Lock is outcome-invariant, not outcome-blind. Outcomes remain relevant for learning and future calibration, but they do not retrospectively redefine who held which substantive right. This distinction reflects accountability research separating process from outcome evaluation (Siegel-Jacobs & Yates, 1996; Lerner & Tetlock, 1999; de Langhe et al., 2011; Pollmann et al., 2014). Proposition 2 shows baseline alignment when excess success-credit retention and blame shifting are removed. Corollary 2 adds a necessary boundary: if residual career downside remains large, distortion can shift from over-delegation to rejection or delay of a positive-value decision. Hence the architecture combines attribution discipline with ex-ante protection for procedurally compliant decisions within approved risk appetite, without shielding misconduct or defective process.

5.4 Human-AI Decisions: Independence as an Information Asset

The Information Lock treats imperfectly correlated human judgments as an organizational information asset rather than asking only whether AI is more accurate than an individual. Proposition 3 shows that blending multiple private judgments with the same imperfect AI signal creates a shared error component; Proposition 4 translates this into fewer effective independent checks. The claim is deliberately bounded: if the blending rule is known and invertible, pre-AI recording adds auditability rather than new statistical content. Proposition 5 restores substantive informational value when separate identification changes incentives to produce private information. This interpretation is consistent with research on algorithm aversion/appreciation, private-information use, and advice timing (Dietvorst et al., 2015; Logg et al., 2019; You et al., 2022; Balakrishnan et al., 2025; Yin et al., 2025).

5.5 Board Governance: Separating Decision Ownership, Delegation Accountability, and Outcome Credit

The model separates decision ownership, delegation accountability, and outcome attribution. Decision ownership identifies who held the effective substantive right; delegation accountability evaluates whether rights were transferred appropriately; outcome attribution links the realized result to the recorded decision structure. This distinction is consistent with established separation between decision management and control and with deliberate allocation of board-management rights (Fama & Jensen, 1983; Useem & Zelleke, 2006). For board-reserved matters, the feasible delegation boundary remains below full transfer; for board-governed executive decisions, full substantive transfer may be permissible. The Ex-Ante Decision Rights Record makes these distinctions observable before results are known and prevents asymmetric narratives in which a superior

claims success while assigning failure to a delegate beyond the rights actually transferred.

5.6 Two Locks, One Ex-Ante Commitment Logic

The two mechanisms belong in one architecture because each preserves a baseline before a later event makes retrospective reconstruction vulnerable to distortion. The Information Lock records designated judgments before common AI influence; the Accountability Lock records effective rights and attribution before outcomes are realized. This is a theoretical integration, not a claim that the two controls always reinforce each other behaviorally. A decision may present only one of the two problems. The contribution is therefore selective: governance should identify which pre-event baseline is economically valuable and apply the corresponding lock rather than impose the full architecture on every decision.

5.7 Contemporary Examples as Motivation, Not Evidence

The UAE Cabinet AI Advisor and IHC's Aiden Insight are used only to motivate the institutional relevance of AI-assisted high-level decision support (Abu Dhabi Media Office, 2024; UAE Government Media Office, 2026). They show that AI can enter consequential decision processes while formal authority remains human, but they do not provide data to estimate the model, reconstruct individual judgment sequences, or measure private career incentives. The paper therefore does not treat either example as causal validation; the theoretical claims rest on the formal model and the cited scholarly literature.

6. Limitations and Future Research

6.1 Model Scope and Endogenous Career Incentives

The formal model remains deliberately parsimonious. Success probability changes linearly with delegated authority, coordination cost is quadratic, and private career payoffs enter through excess credit retention, blame shifting, and a residual outcome-contingent career-loss extension. These choices make the attribution wedge and rejection-substitution boundary transparent but do not exhaust the ways career concerns can operate. Future models can endogenize delegate selection, bargaining over authority, repeated promotion incentives, nonlinear downside exposure, organizational insurance, and the governance process that determines process protection. They can also examine under-delegation caused by fear of losing legitimate status or visibility.

6.2 Information Structure and AI Behavior

The information theorem uses equal-variance human signals, a common AI-advice weight, and baseline independence between AI error and pre-exposure human error. Boundary Result 1 shows that the simplest linear model is perfectly reconstructible when the AI signal and weight are known, while Proposition 5 introduces one specific information-acquisition friction. Real boards may instead exhibit heterogeneous expertise, unknown or unstable AI weights, discrete votes, strategic misreporting, memory effects, correlated source data, and AI systems that adapt to prior human inputs. Extensions should permit heterogeneous weights, endogenous advice seeking, richer signal-acquisition technologies, and AI that behaves as a strategic or agentic participant rather than as a passive information source.

6.3 Institutional Heterogeneity, Board Politics, and Legal Boundaries

Decision rights are embedded in legal and institutional structures that differ across jurisdictions, ownership forms, and sectors. Some matters are statutorily reserved to boards; others are management decisions subject only to oversight. Boards also vote, form coalitions, defer to chairs, and operate under fiduciary and regulatory duties that are not modeled here. The $\bar{\lambda}$ boundary captures the existence of non-delegable rights but not the full political economy of boards. Empirical work should therefore test the model separately in board-reserved and board-governed executive decisions and should distinguish formal delegation documents from effective authority in practice.

6.4 Dynamics, Emergencies, and Empirical Validation

The model is one-shot, whereas many strategic decisions unfold through repeated reviews, staged investments, reversals, crises, and learning. Delay can also dominate the value of additional governance in emergencies. A natural extension is a dynamic model in which the lock can be activated at specific gates rather than at every decision cycle. Empirically, self-protective delegation should be measured from observable rights, timing, and attribution behavior rather than from Likert-style self-reports. Laboratory or field experiments could vary personal downside, delegation rights, AI-advice timing, and pre-commitment rules; archival studies could code delegation records, board papers, decision logs, and subsequent credit/blame narratives. Such designs would test whether the theoretical wedges appear in consequential organizational settings.

7. Conclusion, Implications, and Recommendations

AI-assisted risk governance creates two distinct opportunities for decision architecture to lose value. Before a decision is finalized, a common AI recommendation can make multiple subsequent human judgments less independent and, under information-production frictions, can reduce incentives to develop separately identifiable private assessments. After a risky decision is realized, ambiguous attribution can make delegation privately attractive when a senior decision-maker can retain success credit or shift failure accountability beyond the rights actually transferred.

The Dual-Lock model treats both problems as ex-ante commitment problems. The Information Lock records designated human assessments before common AI exposure and then permits revision. The Accountability Lock fixes effective decision rights and attribution rules before the outcome is known while preserving separate process accountability for the quality of delegation. AI remains an adviser, and legal or organizational decision ownership remains human.

The formal analysis strengthens four claims. First, firm-optimal delegation depends on the decision's stakes and on how transferred authority changes success probability, capacity, and coordination cost. Second, excess success-credit retention and blame shifting create an endogenous attribution wedge that moves executive-optimal delegation away from the firm benchmark; eliminating both excess attribution channels restores baseline alignment only when no additional outcome-contingent private career penalty remains. Third, if residual career downside is sufficiently large, the distortion can shift from over-delegation to rejection of a positive-value decision, which creates a role for ex-ante process-based risk acceptance. Fourth, common exposure to the same imperfect AI signal induces shared error covariance among revised human judgments; when post-AI reports are perfectly invertible the lock adds auditability rather than new statistical information, while an information-acquisition extension shows how separately identifying private assessments can preserve incentives to produce independent information.

The practical implication is selective rather than universal. Boards should preserve an independent pre-AI baseline where distributed human information is valuable and cannot otherwise be reliably reconstructed or where common advice would crowd out independent signal production. They should impose outcome-invariant attribution rules where delegation and career-risk asymmetry create a material incentive problem, while also pre-committing that a procedurally compliant decision within approved risk appetite will not be condemned merely because the stochastic outcome is unfavorable. Routine, reversible, low-stake decisions need not bear the same procedural burden. In this sense, risk governance is not the avoidance of risky decisions. It is the design of decision processes that preserve independent information, allocate decision rights transparently, and distinguish bad luck from bad process.

As AI advisers become more capable, the central governance question is unlikely to be whether humans or machines 'own' the decision in the abstract. It will be whether organizations can preserve the informational diversity and accountability structure needed for human authority to remain meaningful. Before advice and before outcome are the two moments at which that meaning is easiest to protect.

Generative AI and AI-Assisted Technologies Disclosure

OpenAI ChatGPT (GPT-5.6 Sol) was used during manuscript development to assist with structured literature discovery and comparison, formal-model development and stress testing, drafting and restructuring, consistency checks, and language editing. The tool materially assisted the formalization and presentation of the model, including the examination of alternative boundary conditions. The research problem, substantive governance principles, interpretation of the literature, and final decisions about the claims and conclusions were directed by the human author. AI-assisted factual and citation claims were checked against original publications, DOI records, publisher pages, or official institutional sources during manuscript preparation, and AI-assisted wording and mathematical expressions were reviewed and revised. The human author takes full responsibility for the accuracy, originality, citations, formal arguments, and conclusions. No original empirical research data were generated, fabricated, altered, or manipulated by generative AI, and no AI system is listed as an author.

References

Abu Dhabi Media Office. (2024, February 27). Artificial intelligence board observer appointed by International Holding Board of Directors. Retrieved from https://www.mediaoffice.abudhabi/en/economy/artificial-intelligence-board-observer-appointed-by-international-holding-board-of-directors/

Aghion, P., & Tirole, J. (1997). Formal and real authority in organizations. Journal of Political Economy, 105, 1-29. doi:10.1086/262063

Ahdadou, M., Ahdadou, A., Aajly, A., & Tahrouch, M. (2026). Artificial intelligence in corporate boards: A dual-dimensional framework for integration across autonomy and structural levels. Data Science and Management, 9, 100174. doi:10.1016/j.dsm.2025.12.001

Ambrus, A., Baranovskyi, V., & Kolb, A. (2021). A delegation-based theory of expertise. American Economic Journal: Microeconomics, 13, 373-419. doi:10.1257/mic.20190008

Argenton, C., Potters, J., & Yang, Y. (2023). Receiving credit: On delegation and responsibility. European Economic Review, 158, 104522. doi:10.1016/j.euroecorev.2023.104522

Balakrishnan, M., Ferreira, K. J., & Tong, J. (2025). Human-algorithm collaboration with private information: Naive advice-weighting behavior and mitigation. Management Science, 72, 265-284. doi:10.1287/mnsc.2022.03850

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. The Review of Economic Studies, 79, 67-87. doi:10.1093/restud/rdr023

Byford, M. C. (2017). Moral hazard in strategic decision making. International Journal of Industrial Organization, 55, 114-136. doi:10.1016/j.ijindorg.2017.10.001

Central Bank of the United Arab Emirates. (n.d.). Structure and governance of the board. CBUAE Rulebook. Retrieved from https://rulebook.centralbank.ae/en/rulebook/5-structure-and-governance-board

Chari, M. D. R., David, P., Duru, A., & Zhao, Y. (2019). Bowman's risk-return paradox: An agency theory perspective. Journal of Business Research, 95, 357-375. doi:10.1016/j.jbusres.2018.08.010

de Langhe, B., van Osselaer, S. M. J., & Wierenga, B. (2011). The effects of process and outcome accountability on judgment process and performance. Organizational Behavior and Human Decision Processes, 115, 238-252. doi:10.1016/j.obhdp.2011.02.003

Deimen, I., & Szalay, D. (2019). Delegated expertise, authority, and communication. American Economic Review, 109, 1349-1374. doi:10.1257/aer.20161109

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. doi:10.1037/xge0000033

Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. The Journal of Law and Economics, 26, 301-325. doi:10.1086/467037

Gillenkirch, R. M., & Velthuis, L. (2023). Delegated risk-taking, accountability, and outcome bias. Journal of Risk and Uncertainty, 67, 137-161. doi:10.1007/s11166-023-09414-2

Gormley, T. A., & Matsa, D. A. (2016). Playing it safe? Managerial preferences, risk, and agency conflicts. Journal of Financial Economics, 122, 431-455. doi:10.1016/j.jfineco.2016.08.002

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. American Economic Review, 100, 1826-1846. doi:10.1257/aer.100.4.1826

Hill, A. (2015). Does delegation undermine accountability? Experimental evidence on the relationship between blame shifting and control. Journal of Empirical Legal Studies, 12, 311-339. doi:10.1111/jels.12074

Hutzschenreuter, T., Kleindienst, I., & Perkhofer, F. (2026). A question of trust: Dynamics of the perceived uncertainty-trust dilemma with artificial intelligence in strategy process initiation. Managerial and Decision Economics, 47, 944-961. doi:10.1002/mde.70078

Lerner, J. S., & Tetlock, P. E. (1999). Accounting for the effects of accountability. Psychological Bulletin, 125, 255-275. doi:10.1037/0033-2909.125.2.255

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. doi:10.1016/j.obhdp.2018.12.005

OECD. (2023). G20/OECD Principles of Corporate Governance 2023. Paris, France: OECD Publishing. doi:10.1787/ed750b30-en

Peker, C. (2026). Optimal linear aggregation of correlated expert judgments. Economics Letters, 261, 112855. doi:10.1016/j.econlet.2026.112855

Pollmann, M. M. H., Potters, J., & Trautmann, S. T. (2014). Risk taking by agents: The role of ex-ante and ex-post accountability. Economics Letters, 123, 387-390. doi:10.1016/j.econlet.2014.04.004

Schneider, S., & Leyer, M. (2019). Me or information technology? Adoption of artificial intelligence in the delegation of personal strategic decisions. Managerial

and Decision Economics, 40, 223-231. doi:10.1002/mde.2982

Siegel-Jacobs, K., & Yates, J. F. (1996). Effects of procedural and outcome accountability on judgment quality. Organizational Behavior and Human Decision Processes, 65, 1-17. doi:10.1006/obhd.1996.0001

Stein, V., & Wiedemann, A. (2016). Risk governance: Conceptualization, tasks, and research agenda. Journal of Business Economics, 86, 813-836. doi:10.1007/s11573-016-0826-4

Stein, V., Wiedemann, A., & Bouten, C. (2019). Framing risk governance. Management Research Review, 42, 1224-1242. doi:10.1108/MRR-01-2019-0042

Stein, V., Wiedemann, A., Hiebl, M. R. W., & Kraus, S. (2026). Risk governance in a rapidly changing competitive environment: An integrative perspective. Strategic Change, 35, 459-469. doi:10.1002/jsc.70095

UAE Government Media Office. (2026, September 2). Mohammed bin Rashid chairs UAE Cabinet meeting, outlines three priorities for government. Retrieved from https://www.mediaoffice.ae/en/news/2026/september/02-09/mohammed-bin-rashid-chairs-uae-cabinet-meeting

Useem, M., & Zelleke, A. (2006). Oversight and delegation in corporate governance: Deciding what the board should decide. Corporate Governance: An International Review, 14, 2-12. doi:10.1111/j.1467-8683.2006.00479.x

Yin, J., Ngiam, K. Y., Tan, S. S.-L., & Teo, H. H. (2025). Designing AI-based work processes: How the timing of AI advice affects diagnostic decision making. Management Science, 71, 9361-9383. doi:10.1287/mnsc.2022.01454

You, S., Yang, C. L., & Li, X. (2022). Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? Journal of Management Information Systems, 39, 336-365. doi:10.1080/07421222.2022.2063553

Zhao, J., & Wei, X. (2026). Accountable AI in the boardroom: A doctrinal and empirical analysis through corporate reports. Computer Law & Security Review, 61, 106325. doi:10.1016/j.clsr.2026.106325

Zhao, Q., Li, T., Li, S., Pan, Y., Wang, H., & Liao, C. (2026). Losing the hand on the wheel: AI trust, decision delegation, and displacement of responsibility in financial decision-making. Acta Psychologica, 268, 107267. doi:10.1016/j.actpsy.2026.107267