

# Responsible AI Governance in Organizational Decision-Making: A Conceptual Framework of Calibrated Trust, Ethical Acceptability, and Decision Legitimacy.

Mohamed Hamza Bennani <sup>1\*</sup>  and Driss Helmi <sup>2</sup>  and Hayat Bennani <sup>3</sup>

<sup>1</sup> Laboratory of Actionable Management and Organizational Decision-Making (MADO), Department of Economics and Management, Faculty of Economics and Management, Mohammed First University, Oujda 60000, Morocco;

<sup>2</sup> Interdisciplinary Research Laboratory in Economics, Finance and Organizational Management, Department of Finance and Organizational Management, Faculty of Economics and Management, Sidi Mohamed Ben Abdellah University, Fez 30000, Morocco;

<sup>3</sup> Higher Institute of Nursing and Health Technology Professions, Rabat 10000, Morocco.

\* **Corresponding author:** Mohamed.hamza.bennani@ump.ac.ma.

**ABSTRACT:** By functioning in an advisory capacity, AI systems are playing an increasingly central role in organizational decision-making, yet their value cannot be assessed based solely on predictive performance. This article describes the governance-centered conceptual model of how technological governance capability and organizational governance capability jointly shape decision quality and decision legitimacy, both of which are distinct outcomes of AI-augmented decision-making. Our proposal not only considers calibration of trust in AI but also ethical acceptability (ecological validity) as parallel mechanisms rather than a fixed serial sequence. Calibrated trust is whether a decision-maker can count on AI at the level that it claims to perform, and ethical acceptability whether AI-supported decisions are justifiable in fairness, safety, privacy, human oversight, traceability & accountability. Learning culture is suggested as an organizational boundary condition by which the influence of organizational governance capability (with a positive sign) on calibrated trust and ethical acceptability is enhanced, while regulatory institutional pressure is proposed to strengthen (with a positive sign) the link between ethical acceptability and decision legitimacy. The framework also differentiates formative higher-order governance capabilities from evaluative mechanisms at the individual level and clarifies their levels of analysis. An example conceptual application to AI-enabled credit decisions is provided to demonstrate how the framework can be connected to verifiable governance evidence without claiming empirical validation. The model provides a unified explanation of the ways in which AI-governance capabilities can influence sound and institutionally justifiable organizational decisions across levels.

**Keywords:** Artificial intelligence, AI-augmented decision-making, AI governance, Calibrated trust, Ethical acceptability, Decision quality, Decision legitimacy, Multilevel governance.

## I. INTRODUCTION

A characteristic of the transition to this increased algorithmic role for AI in organizations is that, as it progresses from a support analytic function to more active layer of organizational decision-making, predictive models, recommender systems and other algorithmic decision-support tools increasingly influence how alternatives are generated (including options trading), evidence is evaluated (or even simply

what data gets collected) and decisions justified. Never mind the speed of decision-making this transformation involves more than that. This both redistributes cognitive work and decision authority between humans and machines, and poses a governance problem not addressed by performance measures alone: even when an AI-assisted decision leads to a beneficial outcome, is the process trustworthy, justifiable, fair, and accountable to those who implement it or are affected by it? Furthermore, literature on AI augmentation likewise emphasises that sustainable value comes not just from machine autonomy but through the embedding of technical capabilities to human and organisational routines [1-3].

Recent evidence from Moroccan higher education also shows that organizational and social conditions significantly shape AI adoption. Ouzif et al. found that facilitating conditions and social influence are important determinants of AI-adoption intentions among Moroccan university teachers [4]. Bennagem Touati et al. likewise revealed strong positive influences of enabling conditions and social influence on the acceptance of ChatGPT by university instructors [5]. These results underscore the need for a nuanced perspective on AI in organizations and provide an impetus to move from adoption-focused questions of technology development to governance-focused questions of responsible embedding of AI within organizational decision-making processes.

Most of this challenge has been covered in the existing literature along different strands. Technical work centers on reliability, robustness, explainability, and model performance. Behavioral studies focus on trust, reliance, acceptance and aversion to AI algorithms; Responsible-AI governance research deals with the accountability, fairness, mitigation of risk and mechanism for organizational controls which are all designed to ensure compliance with regulations. The streams are closely intertwined but they are theorized separately which means that the extant literature is unable to tell an account that fully explains how governance arrangements translate into decision outcomes within organizations. Recent studies on responsible-AI governance similarly highlight an ongoing chasm between high-level principles and organizational operationalization [6]. The implementation challenge is reflected in international AI-ethics scholarship, which finds commonality across countries on the fundamental principles underlying ethical AI, but significant diversity in their interpretation and implementation [7], as well as competing proposals that show how principles by themselves cannot guarantee ethical AI without institutional mechanisms translating them into accountable organizational practice [8].

This fragmentation matters in particular because qualitatively, decision quality and decision legitimacy are inherently linked but analytically distinct. Even when an AI-supported decision is accurate and timely, it may still lack legitimacy if accountability is unclear, affected stakeholders cannot meaningfully challenge the outcome, or the process lacks procedural or institutional credibility. An open, procedurally careful process will also produce faulty predictions or a failure to be consistent with decisions. As such, technical effectiveness and social or institutional justification should therefore be treated as separate but connected outcomes. Thus, we ask: How might technological and organizational governance capabilities influence the quality and legitimacy of AI-augmented organizational decisions, by which mechanisms, and under what organizational and institutional contingencies? We propose a multilevel conceptual model, wherein technological governance capability and organizational governance capability serve as formative second-order antecedents. They are hypothesized to have effects through two parallel mechanisms at the individual level: calibrated trust in AI and their ethical acceptability. Calibrated trust concerns whether decision-makers align their reliance on AI with its demonstrated capabilities and limitations, whereas ethical acceptability concerns whether the use of AI can be justified in a particular decision context. We treat decision quality and decision legitimacy as distinct outcomes, specify the boundary conditions on selected relationships using learning culture and regulatory institutional pressure, respectively.

The contribution is threefold. First, the framework distinguishes decision quality from decision legitimacy and therefore refuses to imply that AI-supported decisions hold institutional or social ground simply because they are technically sound. This leads to a mechanistic account of governance, articulating a behavioral from a normative pathway where the former proceeds through calibrated trust and the latter through ethical acceptability. Third, it anchors these mechanisms in a multilevel model that connects organizational governance abilities and individual assessments to decisions and institutional constraints. An

illustrative example of how the constructs can be connected with observational evidence in AI-assisted credit decision-making (but without treating this example as empirical validation) is provided.

The rest of the article is structured as follows. Conceptual-Development Approach (Section 2) The theoretical foundations are given in Section 3. The fourth section describes the architecture of governance-centered model and constructs. Section 5 develops the theoretical implications. Section 6 contains an example conceptual application to credit decisions assisted by AI. In section 7, we discuss theoretical contributions, managerial implications, limitations and empirical research agenda.

## II. CONCEPTUAL DEVELOPMENT APPROACH

The article does not take a systematic literature review approach, but is instead an integrative, theory-building synthesis. This is not intended to be an exhaustive list of the AI-governance literature but rather to include theoretically relevant work that helps to explain how governance capabilities are encoded through normative projections into decisions augmented by AI. The literature base was grouped along six inter-related streams: (1) AI augmentation and socio-technical decision making; (2) responsible Artificial Intelligence governance; (3) trust in automation and AI; (4) Organizational legitimacy, (5) organizational learning and psychological safety; and (6) institutional and regulatory governance. We relied on seminal works to anchor well-established theoretical frameworks, and drew upon more recent empirical studies that reflect state-of-the-art developments in AI governance and organizational decision-making.

Sources were included if they provided a direct contribution to at least one of three tasks: the definition of focal construct, an explanation of a proposed relationship or delineation of a boundary condition. Unlike traditional approaches to summarize studies in a linear fashion, the integrative synthesis noted that dimensions emerged repeatedly and subsequently ordered them as per their role in the conceptual model. Reliability, robustness, explainability, auditability and workflow integration were synthesized as proposed dimensions of technological governance capability. Synthesis of Data governance, process formalization, feedback, master decision making issue escalation & accountability were synthesized into Organizational Governance capability dimensions. The calibrated-trust mechanism drew from research on trust in automation and AI, while the constructs of ethical acceptability and decision legitimacy were influenced by responsible-AI and legitimacy research.

There are four stages in the conceptual-development process. In the first stage, recurring governance dimensions identified through the integrative synthesis were conceptually grouped into higher-order antecedent constructs. The second was separating behavioral and normative mechanisms to ensure trust, ethics, and transparency were not interchangeable. Third, outcome space in terms of isolating decision quality and decision legitimacy. Finally, drawing on literature related to the organizational and institutional contexts of governance, we introduce learning culture and regulatory institutional pressure as boundary conditions reflecting important differences in how effective or externally credible certain governance practices might be across those various contexts. The behavioral relationships that were found served as the basis of propositions that define the following aspects: predictor, outcome, expected direction and level of analysis to which they are relevant. As an integrative conceptual synthesis rather than a systematic review, this framework should thus be viewed as theoretically deduced model to be tested empirically in the future.

## III. THEORETICAL FOUNDATIONS

### 1. AI AUGMENTATION AS A SOCIO-TECHNICAL GOVERNANCE PROBLEM

We conceptualize AI-augmented decision-making as a socio-technical arrangement, rather than an isolated predictive artifact. The quality and defensibility of the decisions ultimately produced depend on how AI capabilities are combined with human judgment and organizational governance. AI can shift which individuals offer alternatives, how evidence is assessed, and who has final say [2]. The automation-augmentation paradox has the same implication that AI may substitute for some human judgment and augment other tasks, making allocation of task to UI and human involvement through managing content generation a central design choice [1].

Accordingly, in alignment with this socio-technical perspective, action-design research shows that AI outputs result in organizational value only if they are embedded into decision processes and serviced by explanation, evaluation and human-action mechanisms; instead of just producing them as a standalone output [3]. Responsible-AI governance research too stresses the translation of principles into organizational practices [6]. Hence, we separate technological governance capability from organizational governance capability. The content of the first refers to technical properties and workflow integration that render AI behavior governable, while the second refers to organizational routines, accountability structures and controls through which AI-supported decisions are governed.

## 2. TRUST, TRUSTWORTHINESS, AND CALIBRATED RELIANCE

Since AI may be used by decision-makers in the collaboration process, we deem it important to address how trust affects the use of AI, and whether or not decision-makers will trust recommendations made by AI (to what extent). Trust should nevertheless be distinguished from trustworthiness and observable reliance behavior. AI trustworthiness is the system and governance properties which can warrant a user to rely on AI systems; on the other hand, perceived trustworthiness expresses users' assessment of those properties. Trust is a belief in how much we may depend on AI-supported recommendations in the face of uncertainty, whereas reliance reflects an observable indication of whether a system based upon our trust will be used, weighted or overridden. These differences align with work studying trust in automation and AI [9-11].

It is not maximal trust, but calibrated and reliable trust. This creates overtrust, which leads to automation bias while undertrust causes users to reject a useful system. Calibration thus entails the matching of belief-driven reliance behaviours with actual observed performance, uncertainty and limits of capability within the system. Cues about competence and constraints are offered by reliability, robustness, explanation quality, auditability, communication of uncertainty and the consistency of workflows [12, 10]. Organizational arrangements are important too: data governance processes, feedback loops, opportunities for escalation and clear accountability mechanisms affect what users read as whether reliance is warranted. What calibrated trust is in the context of this framework therefore boils down to an evaluative mechanism at the individual-level governing the quality of decision-making affected by peoples' governance capabilities.

## 3. ETHICAL ACCEPTABILITY AND RESPONSIBLE AI USE

Trust is conceptually distinct from ethical acceptability. For instance, an AI system could meet predictive competence criteria but be unacceptable to a decision-maker due to unfairness in the training data, lack of accountability, and/or limited human oversight or contestability. On the flip side, an organization could build new system but with ethically favourable restrictions before people are ready to depend and act based on what its predicting. This means that ethical acceptability deals with different question than predictive competence: is the application of AI in this case ethically justifiable?

Ethical acceptability within this framework means determining whether the deployment of AI is justifiable with respect to issue areas such as fairness, safety, privacy, human oversight, traceability and auditability. It binds responsible-AI principles to organizational decision processes, where it aligns with those risk-based governance frameworks mandating organization to govern, map, measure and manage AI risks throughout the lifecycle [13, 14]. Since calibrating trust and ethical acceptability address two distinct evaluative questions, they are modeled as independent factors rather than as a linear chain of causation. Serial or reciprocal relationships remain theoretically plausible and are an interesting alternative specification for future longitudinal research. Contextual evidence located in applied AI settings again reveals that ethical acceptability materializes through concerns surrounding privacy, algorithmic bias, transparency on the grounds of a formal ethics training and responsibility and accountability [15, 16]. These studies are founded in education; however, they demonstrated how operational and governance practices highlight ethical principles.

## 4. DECISION QUALITY AND DECISION LEGITIMACY

The quality of decisions is constructed instrumentally, that is, based upon the extent to which an AI-augmented decision delivers on its stated goals. It is currently reflected mostly in accuracy, consistency,

timeliness, effectiveness. In contrast, decision legitimacy refers to whether or not the decision, and the process surrounding how it was made, is regarded as lawful, appropriate justification and acceptable in accordance with applicable procedural, social and institutional standards. Thus, legitimacy theory differentiates between a technical performance and social-institutional assessments of the appropriateness of actions [17-20].

This difference is particularly salient when it comes to AI contexts, since stakeholders might assess a decision not just based on whether it was made properly but whether accountability was maintained, the reasoning was understandable, and anyone affected has real opportunities for review or recourse. Hence, a decision can be quality but not legitimate or procedural legitimacy but poor performance. A Meta-Analyzed Understanding of Traceability as a Governance Dimension to Performance, Legitimacy, or Both (2019) To mitigate construct overlap, traceability is framed as a governance and evidentiary mechanism supporting both performance evaluation and legitimacy assessment rather than being constitutive of outcome.

**Table 1.** Core constructs, levels of analysis, and measurement logic.

| Construct                            | Definition in the model                                                                                                                  | Analytical role   | Primary level                  | Measurement logic               |
|--------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|-------------------|--------------------------------|---------------------------------|
| Technological governance capability  | Capability constituted by reliability/robustness, useful explainability, auditability, and workflow integration.                         | Antecedent        | Organizational / system        | Formative higher-order          |
| Organizational governance capability | Capability constituted by data governance, process formalization, feedback, escalation, and explicit accountability.                     | Antecedent        | Organizational                 | Formative higher-order          |
| Calibrated trust in AI               | Willingness to rely on AI in proportion to perceived competence, limits, and integrity.                                                  | Parallel mediator | Individual                     | Reflective perceptual mechanism |
| Ethical acceptability                | Perceived defensibility of AI use in terms of fairness, safety, privacy, oversight, traceability, and responsibility.                    | Parallel mediator | Individual                     | Reflective evaluative mechanism |
| Decision quality                     | Instrumental performance of the decision in terms of accuracy, consistency, timeliness, and effectiveness.                               | Outcome           | Decision                       | Outcome construct               |
| Decision legitimacy                  | Perceived propriety and justifiability of the decision under relevant procedural, social, and institutional expectations.                | Outcome           | Decision / stakeholder         | Outcome construct               |
| Learning culture                     | Psychological safety and routines supporting questioning, error reporting, review, and corrective learning.                              | Moderator         | Organizational / team          | Contextual construct            |
| Regulatory institutional pressure    | Clear, monitored, and consequential external expectations for AI governance, varying by jurisdiction, sector, and enforcement intensity. | Moderator         | Institutional / jurisdictional | Contextual construct            |

## IV. GOVERNANCE-CENTERED CONCEPTUAL MODEL

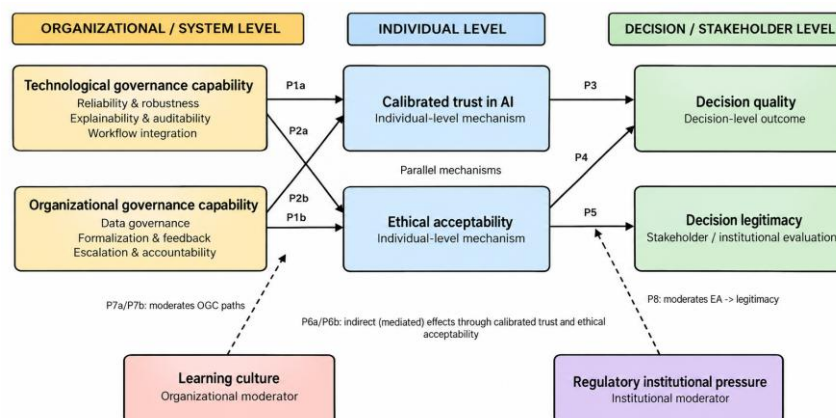
### 1. MODEL ARCHITECTURE AND LEVELS OF ANALYSIS

We analyze the model of two formative governance capabilities, two parallel mechanisms at the individual level, two separated results and moderate by context. The Technological Governance capability describes the ability to govern AI behaviour from an engineering standpoint with reliability & robustness, decision-relevant explainability, and auditability, but also the right level of integration into a structured workflow. Organizational governance capability describes the structures and routines aimed at governing

decisions based on AI and includes data governance, process formalization, feedback, escalation and accountability. They are both understood as formative higher order constructs since their respective dimensions are not interchangeable and jointly capture this overall capability. Thus, an AI system might be reliable but not explainable and an organization strong on data governance but weak on accountability; the higher-order capability is thus defined through a configuration of these dimensions rather than uniformly by them.

The model is explicitly multilevel. The capacities for technical and organizational governance primarily reside at the organization or system level, along with learning culture. Both calibrated trust and ethical acceptability are assessments at the level of each individual. Decision quality is assessed at the decision level; decision legitimacy is evaluated by stakeholders and other institutional audiences. Regulatory institutional pressure exists on the institutional (or jurisdictional) level. At the cross-level, those organizational governance arrangements determine how individual decision-makers assess whether AI is trustworthy enough to decide and ethically defensible enough to deploy; in turn, these individual assessments influence decision-making outcomes, while institutional pressures shape how practices deemed ethically-defensible benefit from legitimacy.

Calibrated trust and ethical acceptability are posited to be parallel, not serial, mechanisms. Calibrated trust: This considers whether a decision-maker can trust AI based on the demonstrated ability and limitations of AI Artificial Intelligence Ethics: On the other hand, ethical acceptability is around establishing defensible reasons to use the qualified output from AI in the relevant context. The model makes no assumptions about the sequencing of either evaluation. It is possible that reciprocal influence may develop over longer periods of time, but including it within the baseline framework would mask the independent effects that the model is meant to parse out. Future longitudinal studies may thus compare the proposed parallel specification with serial and reciprocal alternatives. A learning culture provides the boundary condition in which the organization operates (for example, Formal governance arrangements are more strongly associated with calibrated trust and ethical acceptability, if employees can challenge AI outputs and anomalies, seek to speak truth to power without personal or organizational sanction [21]. Regulatory institutional pressure serves as an externality that drives expectations around fairness, documentation, oversight, and auditability into the forefront of public discourse. It does not guarantee legitimacy on its own, but it amplifies the connection between ethical AI use and stakeholder perceptions about decision legitimacy.



**FIGURE 1.** Revised governance-centered conceptual model of AI-augmented decision-making. Solid arrows represent proposed direct relationships; dashed arrows represent moderation. Calibrated trust and ethical acceptability are parallel mechanisms.

## 2. THEORETICAL DISTINCTIVENESS RELATIVE TO ADJACENT FRAMEWORKS

This framework does not seek to supplant existing theories of responsible-AI governance, human-AI decision-making, trust, or legitimacy. Its integrative contribution lies in specifying how governance

capabilities can lead to two distinct decision outcomes through separate mediating mechanisms operating across different levels of analysis. Responsible-AI governance research identifies principles, risks, and governance practices; human-AI decision research clarifies the allocation of judgment and authority; trust research explains reliance; and legitimacy research explains judgments of propriety. The present framework organizes these insights within a common causal architecture focused on organizational decision outcomes.

**Table 2.** Positioning the framework relative to adjacent research streams.

| Research stream                             | Primary focus                                     | What it explains well                                              | Gap addressed by the present framework                                                                   |
|---------------------------------------------|---------------------------------------------------|--------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Responsible AI governance (e.g., [6], [13]) | Principles, risks, controls, governance practices | What organizations should govern and institutionalize              | How governance capabilities translate into decision quality and legitimacy through distinct mechanisms   |
| Human-AI decision-making (e.g., [1], [2])   | Task allocation, augmentation, authority          | How humans and AI interact in organizational decisions             | Adds decision legitimacy as a distinct outcome alongside instrumental performance                        |
| Trust in AI / automation (e.g., [9], [10])  | Trust, reliance, calibration                      | Why users rely on or reject AI                                     | Embeds calibrated trust within organizational governance and distinguishes it from ethical acceptability |
| Legitimacy theory (e.g., [17], [19])        | Propriety, acceptance, institutional expectations | Why organizational actions are judged appropriate or inappropriate | Links legitimacy to AI governance while keeping it distinct from decision quality                        |

## V. THEORETICAL PROPOSITIONS

### 1. GOVERNANCE CAPABILITIES AND CALIBRATED TRUST

Competence A consistent and robust AI output signals capability auditability and interpretability provide the means to inspect the basis and limitations of system behavior. It also clarifies when and how we should use AI recommendations by specifying appropriate workflow integration. In union, these dimensions mitigate uncertainty about when one should depend on specific authorities. Stronger technological governance should hence enable better-calibrated trust as reliability, robustness, explainability, auditability and workflow integration is a source of evidence for users about system ability and limits [12, 10].

- P1a. Technological governance capability is positively associated with calibrated trust in AI.

Meanwhile, the wider organizational environment around the technology helps in shaping trust. Two, data governance instills trust in the validity, lineage, and authorized use of inputs. Formal review and feedback mechanisms signal that model outputs are not taken as *fait accompli*. Who owns an AI-backed decision, and what happens when the system fails or uncertainties are high is made clear by escalation procedures and explicit accountability. Such arrangements mitigate uncertainty about suitable use and failure response, which in turn enables more effective trust calibration.

- P1b. Organizational governance capability is positively associated with calibrated trust in AI.

### 2. GOVERNANCE CAPABILITIES AND ETHICAL ACCEPTABILITY

Potential harms are made more visible and governable to the beneficiaries of these capabilities through reliable performance, decision-relevant explanations, auditability, and integration in relevant workflows (exhibit 4). While technical performance does not itself establish ethical defensibility, this type of technical governance capability indicates something that could be assessed in terms of fairness, safety and adequacy of human involvement.

- P2a. Technological governance capability is positively associated with the perceived ethical acceptability of AI-supported decision-making.

Further, improvization in organizational governance ability without negating ethical acceptability. Concrete rules on data, channels to provide feedback, escalation mechanisms and accountability for implementing these embed the lofty sentiments of fairness, privacy, oversight and responsibility in organizational practice. Research on responsible-AI governance also focuses on turning principles into routines or controls [6].

- P2b. Organizational governance capability is positively associated with the perceived ethical acceptability of AI-supported decision-making.

### 3. CALIBRATED TRUST AND DECISION QUALITY

It defines how much trust is put into AI recommendations versus human judgement. The evidence-based way of being right-sized allows decisionmakers to deploy AI wherever it has demonstrated a relative comparative advantage, while still retaining human judgement for circumstances in which uncertainty, novelty or disclosure of known flaws would benefit from some form of intervention. That is, the relationship that we care about here is not 'more trust is better', but mid-calibration: blind faith creates automation bias [22], while chronic distrust would inhibit AI from making beneficial contributions [9, 10].

- P3. Better-calibrated trust in AI that is, closer alignment between reliance and demonstrated system capabilities and limitations is positively associated with decision quality in AI-augmented organizational decision-making.

### 4. ETHICAL ACCEPTABILITY AND DECISION OUTCOMES

Ethics acceptability as a governance practice, by making issues of data provenance, fairness, oversight and responsibility and system limits visible and executable can also reinforce decision quality. These safeguards minimize avoidable errors arising from improper use of data, ineffective monitoring, or inadequate review. Consequently, ethical acceptability can serve both an instrumental function (e.g. aiding decision processes) and a normative function (enabling actively defensible use of AI).

- P4. Higher ethical acceptability of AI-supported decision-making is positively associated with decision quality.

The more direct relation to decision legitimacy is on ETHICAL ACCEPTABILITY. Stakeholders assess the extent to which a relevant norm is upheld by a decision process and whether accountability, explanation and meaningful contestability are preserved [17, 18]. Such qualities can make an AI-backed choice at least procedurally defensible even if it goes against the interests of specific stakeholders.

- P5. Higher ethical acceptability of AI-supported decision-making is positively associated with decision legitimacy.

### 5. PARALLEL MEDIATION

The framework proposes that governance capabilities also influence decision outcomes indirectly. Calibrated trust transmits the effects of governance capabilities to decision quality by shaping the appropriateness of users' reliance on AI. Ethical acceptability transmits governance effects to both decision quality and decision legitimacy by shaping whether AI use is perceived as sufficiently governed and defensible. The two mediators are specified as parallel rather than sequential mechanisms. Future empirical studies may compare this parallel mediation structure with alternative serial or reciprocal specifications and may also test residual direct effects.

- P6a. Calibrated trust mediates the positive relationships between technological and organizational governance capabilities and decision quality.
- P6b. Ethical acceptability mediates the positive relationships between technological and organizational governance capabilities and decision quality and decision legitimacy.

### 6. LEARNING CULTURE AS AN ORGANIZATIONAL BOUNDARY CONDITION

Organizational climate that demystifies questioning and encourages error reporting, review and corrective learning creates a greater prospect of formal governance affecting employee evaluations and behavior. Psychological safety allows employees to challenge AI outputs, record outliers and trigger

escalation processes without risk of interpersonal or organizational backlash [21]. In such conditions, governance arrangements are likely to be felt as effective and responsive rather than symbolic.

- P7a. Learning culture strengthens the positive relationship between organizational governance capability and calibrated trust in AI, such that the relationship is stronger when learning culture is high.
- P7b. Learning culture strengthens the positive relationship between organizational governance capability and ethical acceptability, such that the relationship is stronger when learning culture is high.

#### 7. REGULATORY INSTITUTIONAL PRESSURE AS AN INSTITUTIONAL BOUNDARY CONDITION

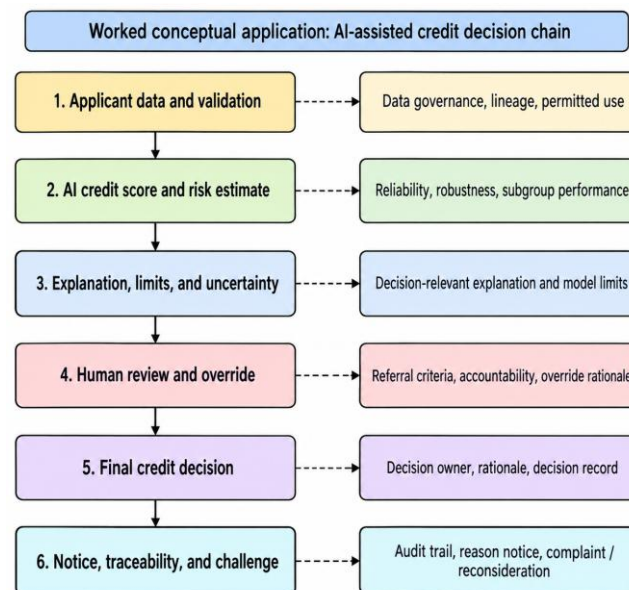
Clarity, monitoring, enforceability and consequences vary across jurisdictions or sectors; regulatory institutional pressure. However, stronger pressure can raise the external legitimacy of internal safeguards by making expectations about documentation, fairness, oversight and accountability more salient and impactful. Regulatory pressure does not equate to legitimacy, but it can reinforce the linkage between ethically defensible organizational practices and stakeholder-led assessments about decision legitimacy.

- P8. Regulatory institutional pressure strengthens the positive relationship between ethical acceptability and decision legitimacy.

## VI. WORKED CONCEPTUAL APPLICATION: AI-ASSISTED CREDIT DECISION-MAKING

### 1. WHY CREDIT DECISIONS PROVIDE AN INFORMATIVE APPLICATION

The domain of AI-assisted credit decision-making offers an informative illustrative setting where technical performance and fairness, explainability, and procedural justification are all intertwined within the same decision process. Credit scoring has the potential to increase risk discrimination, stability, and processing time (which can help lenders), yet it could be perpetuating historical bias using suspect data or masking the reasons for an adverse action decision. Research along these lines therefore now examines performance, fairness, explainability and governance in lending and credit-scoring contexts together [23-27].



Illustrative only: the diagram does not constitute empirical validation.

**FIGURE 2.** Worked conceptual application of the governance framework to an AI-assisted credit decision. The application is illustrative and does not constitute empirical validation.

In this institutional setting, governance is especially salient. Artificial intelligence systems applied to assess the creditworthiness of natural persons or set a credit score generally fall under the high-risk category

according to the European Union AI Act, subject to exceptions [28]. This article is not intended as a legal analysis of credit regulation. Instead, it is a hypothetical worked example demonstrating how constructs from this proposal could be fitted to an important decision-making process and linked with observable behavioural governance evidence.

## 2. MAPPING THE FRAMEWORK TO THE DECISION CHAIN

For illustration, focus on a retail lender which uses an AI model based on information about an applicant to predict probability of default and aid in the decisions whether to approve, reject, price or refer. The governance question is not just whether the model is a good predictor on average; it is also about whether application of the end-to-end decision process yields an outcome that can be made sound, accountable and explainable both for the institution (the one deploying this process) as well as for the applicant. As a result, governance evidence is required at each stage of the decision chain, from data collection and model development to decision-making, pricing, explanation, and ongoing monitoring. Organizational governance at data intake relates to provenance, quality, permissions and access rights to sensitive or proxy variables. Technological governance deals with calibration, robustness, drift and subgroup performance in model scoring. The institution wants decision-relevant reasons at the explanation stage with model limitation clearly communicated. The efficacy of human review relies on clear referral thresholds, override guidance, reviewer competence and documentation requirements. The final decision necessitates clear responsibility and a rationale that is able to be reconstructed, while governance after-the-fact requires traceability, notification, complaints channels, along with a substantive process for reconsideration.

**Table 3.** Mapping the conceptual model to AI-assisted credit decisions.

| Decision stage       | Governance question                                                                | Observable evidence                                                                           | Primary construct(s)                                                            |
|----------------------|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------|
| Data intake          | Are the data valid, relevant, traceable, and appropriately governed?               | Data dictionary; lineage; missingness report; prohibited/proxy variable review; access rights | Organizational governance capability                                            |
| AI scoring           | Is the model reliable across time and relevant applicant groups?                   | Calibration; error rates; stability/drift tests; subgroup performance; robustness tests       | Technological governance capability                                             |
| Explanation          | Can the recommendation and its limitations be understood where action is required? | Reason codes; local explanation; uncertainty/confidence; documented model limits              | Technological governance capability / calibrated trust in AI                    |
| Human review         | When should a human question, refer, or override the model?                        | Referral thresholds; override criteria; reviewer competence; override reasons and rates       | Calibrated trust / ethical acceptability                                        |
| Final decision       | Who owns the decision and can its basis be reconstructed?                          | Decision owner; approval log; model version; data snapshot; rationale                         | Organizational governance capability / ethical acceptability / decision quality |
| Notice and challenge | Can the affected person understand and meaningfully contest the outcome?           | Adverse-action reason; complaint route; response time; reconsideration procedure              | Decision legitimacy / ethical acceptability                                     |

## 3. FROM CONSTRUCTS TO POTENTIAL EMPIRICAL MEASURES

These constructs are only helpful for future empirical research if they can be tied to real evidence. Although perceptual measures are still essential for calibrated trust, ethical acceptability and legitimacy, they

should be accompanied by the behavioral, operational and documentary evidence. For example, surveys may show approval that credit officers understand and trust the AI system, while override logs can help gauge actual reliance v. those impressions. The same applies to the perceptions of procedural fairness or legitimacy, which can be cross related with complaints patterns, reconsideration outcomes and audit or regulatory findings.

**Table 4.** Construct-to-proposition-to-measure mapping for future empirical research.

| Construct                            | RP                             | OI/IT                                   | Potential empirical measure                                                                                                 | Evidence source                                                                            |
|--------------------------------------|--------------------------------|-----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| Technological governance capability  | P1a, P2a, P6a, P6b             | Objective / documentary indicators      | Reliability and calibration metrics; robustness results; explanation usability; auditability; workflow-integration measures | Validation reports; monitoring logs; model cards; audit records; workflow documentation    |
| Organizational governance capability | P1b, P2b, P6a, P6b, P7a, P7b   | Documentary / organizational indicators | Maturity of data-governance controls; feedback and escalation routines; role clarity; accountability measures               | Policies and governance manuals; escalation logs; role matrices; accountability records    |
| Calibrated trust in AI               | P1a, P1b, P3, P6a, P7a         | Perceptual + behavioral indicators      | Validated trust/calibration scales; appropriate-reliance scenarios; recommendation weighting; override rates                | Decision-maker surveys; experiments or scenarios; decision and override logs               |
| Ethical acceptability                | P2a, P2b, P4, P5, P6b, P7b, P8 | Perceptual / evaluative indicators      | Fairness, safety, privacy, oversight, traceability, and accountability judgments                                            | Surveys; ethics assessments; review records; governance documentation                      |
| Decision quality                     | P3, P4, P6a, P6b               | Objective operational indicators        | Accuracy; error rate; consistency; timeliness; downstream decision effectiveness                                            | Decision datasets; performance dashboards; operational records                             |
| Decision legitimacy                  | P5, P6b, P8                    | Perceptual + institutional indicators   | Procedural fairness and justifiability scores; complaint and reconsideration rates; audit/regulatory findings               | Stakeholder surveys; complaint files; reconsideration records; audit or regulatory reports |
| Learning culture                     | P7a, P7b                       | Perceptual / organizational indicators  | Psychological safety; error-reporting behavior; review and corrective-learning routines                                     | Employee surveys; incident reports; review and remediation logs                            |
| Regulatory institutional pressure    | P8                             | Regulatory / documentary indicators     | Regulatory clarity; enforcement intensity; audit requirements; jurisdictional maturity; sector-specific obligations         | Statutes and regulations; regulator guidance; enforcement actions; audit records           |

Note: Related proposition(s) = RPs, Observable indicator / indicator type = OI/IT

This distinction enables empirical research to evaluate decision quality separately from decision legitimacy. If affected applicants are not provided with adequate reasons for rejection or an effective mechanism for review, a lender may improve predictive accuracy and processing efficiency while still facing a legitimacy deficit. Conversely, a transparent and contestable decision process may still rely on a poorly

performing model. Performance and legitimacy should therefore be evaluated separately rather than combined into a single overall score.

## VII. DISCUSSION AND CONCLUSION

### 1. THEORETICAL CONTRIBUTIONS

Our first input is that "decision quality" is separate from decision legitimacy. The AI governance is sometimes discussed as if technically sound systems must produce socially acceptable outcomes. The framework rejects that equivalence. Decision quality addresses the instrumental effectiveness of the organizational decision, whereas decision legitimacy pertains to whether or not the process and outcome can be justified by relevant social, procedural, and institutional expectations. This distinction allows researchers to investigate situations where performance and legitimacy may diverge.

The second contribution is a governance account predicated on the mechanism. However, it is not purely a question of desirable practices, but rather how decision-makers assess and use AI; thus, if relevant to the study on technological and organizational governance capabilities. In particular calibrated trust characterises a behavioural route to appropriate reliance, whilst ethical acceptability describes a normative pathway to defensible use. Considering these mechanisms as parallel sets avoids imposing a universal causal ordering, which is very hard to justify theoretically and that clarifies how different decision outcomes can be brought about by the same governance arrangement but through different mechanisms.

The third input is the multilevel construction of governance. Individual-centered evaluations are influenced by organizational capabilities and learning culture, which affect decision-level outcomes; while regulatory institutional pressure takes place at an even broader institutional level. It is this cross-level architecture that demonstrates why AI governance cannot be categorized as either property of the technical system level or at best disposable attitude at the individual user level. It further pinpoints boundary conditions that might account for the divergence of effects across organizations and jurisdictions when they have similar governance arrangements.

Another logically related implication relates to construct architecture. The framework defines technological and organizational governance capabilities as formative higher order constructs, because they are heterogeneous, noninterchangeable, and jointly constitutive. This specification makes the future measurement problem salient and avoids assuming that the various governance dimensions such as reliability, explainability, accountability must covary as reflective indicators of a single latent variable.

### 2. MANAGERIAL IMPLICATIONS

The framework suggests for managers that governance needs to be at the core of decision chain and not just documentation post deployment. Model validation, data lineage, explanation artifacts, override rules, fairness testing, escalation paths and audit trails are direct controls through which users determine if the AI system can be responsibly used and trusted.

It should be targeted but not symbolic human oversight. Making manual sign-off a requirement for every AI-facilitated recommendation could delay decisions without enhancing accountability. Instead of threshold criteria, organizations should specify the circumstances (high uncertainty, ethically sensitive, novelty, potentially harmful) in which human review is warranted. A similar principle applies to transparency: the goal is not maximal disclosure, but rather information which is relevant for decision-making that facilitates effective action, accountability and meaningful challenge.

### 3. LIMITATIONS AND EMPIRICAL RESEARCH AGENDA

The article remains conceptual. The credit application does not depict an empirical case study but rather a worked conceptual example, and no one of the hypothesized causal chain is claimed to be based on primary organizational data. This is an integrative literature synthesis instead of a systematic or exhaustive synthesis, so we cannot claim to cover all relevant research. Future work will therefore strengthen the measurement models and structural relationships across organizations, sectors, and jurisdictions.

There are three stages an empirical program can follow in the future. The first one is to develop and validate measurement models for the formative higher-order governance capabilities as well as those of calibrated trust, ethical acceptability, decision quality, decision legitimacy, learning culture and regulatory institutional pressure. Second, the identified parallel mediation structure could be contrasted with alternative hypothetical models that included serial and reciprocal associations among calibrated trust and ethical acceptability. Longitudinal designs are particularly useful, because trust governance maturity, incidents, remediation and regulatory expectations changes over time. Third, multilevel (multisite or organizational) designs should conceptualize individuals as nested in organizations and organizations nested in institutional environments (with P7 and P8 tested as organizational- and institutional moderation effects).

To further mitigate common-method bias, perceptual measures can be triangulated with objective and documentary evidence (e.g., override logs, decision latency, model-drift events, changes in data composition, fairness metrics, audit findings, complaint patterns, and reconsideration outcomes [13, 14, 23-26]) which further allows for evaluation across complementary perspectives. Instead of using a simple unidimensional construct, develop operationalizations of institutional pressure with multiple attributes, such as jurisdiction, sectoral focus and type of regulation; clarity and specificity of the regulations (regulatory ambiguity); regulatory enforcement intensity; nature of audit requirements; and institutional maturity.

This leads the framework to move away from an evaluation of the AI-augmented decision-making only against the technical artifact, to a consideration about how it fits into the overall governance system. Properly governed AI should allow organizations to know when they can depend on the output, when it is appropriate to challenge an AI output, what documentation ensures accountability and how decisions are made so that those impacted by them can understand and contest important ones. The core theoretical assertion thus does not contend that bits of governance are inherently better but rather that technological and organizational governance capabilities enable the quality of outcomes and decision legitimacy through different behavioral and normative mechanisms across various levels of analysis.

### Conflicts of Interest

The authors declare no conflicts of interest.

### Author Contributions

Conceptualization, M.H.B. and D.H.; methodology, M.H.B. and D.H.; validation, D.H.; formal analysis, M.H.B.; investigation, M.H.B. and D.H.; resources, D.H.; data curation, M.H.B.; writing—original draft preparation, M.H.B.; writing—review and editing, D.H.; visualization, M.H.B.; supervision, D.H.; project administration, D.H. All authors have read and agreed to the published version of the manuscript.

### Funding Statement

This study has no external funding.

### Data Availability Statement

Data are available from the authors upon request.

### Acknowledgments

Not applicable.

### REFERENCES

1. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, 46(1), 192–210.
2. Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4), 66–83.
3. Herath Pathirannehelage, S., Shrestha, Y. R., & von Krogh, G. (2025). Design principles for artificial intelligence-augmented decision making: An action design research study. *European Journal of Information Systems*, 34(2), 207–229.

4. Ouzif, H., El Boukhari, H., El Achabi, M., Chahbouni, O., Es-Sanoun, M., Mahouat, N., & Bakkali, S. (2025). Analyzing the adoption of artificial intelligence by Moroccan university teachers: Key insights and implications from the UTAUT model. *Edelweiss Applied Science and Technology*, 9(4), 2722–2732.
5. Bennagem Touati, A., Hadj-Ali, A., Ouzif, H., Benlamrabet, O., & Es-Sanoun, M. (2025). From hesitation to integration: A UTAUT model analysis of ChatGPT adoption in Moroccan universities. *Edelweiss Applied Science and Technology*, 9(5), 2980–2995.
6. Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), 101885.
7. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1, 389–399.
8. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507.
9. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
10. Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2), 627–660.
11. Lahusen, C., Maggetti, M., & Slavkovik, M. (2024). Trust, trustworthiness and AI governance. *Scientific Reports*, 14, 20752.
12. Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11, 1568.
13. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology.
14. Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile* (NIST AI 600-1). National Institute of Standards and Technology.
15. Makhmudova, D. M., Sharipova, X. R., Hojiyev, N. K., Ergashev, A. E., Satvaldieva, Y. Kh., Mamatkulova, K. U., & Khudoynazarov, E. M. (2025). Ethical and legal challenges of implementing AI in science and math education in Central Asia. *Qubahan Academic Journal*, 5(3), 294–326.
16. Abisheva, C., Koldasbaeva, Z., Nossiyeva, N., Irgebaeva, N., Aipova, A., Doldinova, S., Smoilov, S., Aykenova, R., Umirzakova, L., & Idrissova, M. (2024). Formation of ethical competences for AI use in English foreign language teachings. *Qubahan Academic Journal*, 4(4), 191–205.
17. Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20(3), 571–610.
18. Bitektine, A. (2011). Toward a theory of social judgments of organizations: The case of legitimacy, reputation, and status. *Academy of Management Review*, 36(1), 151–179.
19. Bitektine, A., & Haack, P. (2015). The “macro” and the “micro” of legitimacy: Toward a multilevel theory of the legitimacy process. *Academy of Management Review*, 40(1), 49–75.
20. Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of Management Annals*, 11(1), 451–478.
21. Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2), 350–383.
22. Wen, Y., Wang, J., & Chen, X. (2025). Trust and AI weight: Human-AI collaboration in organizational management decision-making. *Frontiers in Organizational Psychology*, 3, 1419403.
23. Babaei, G., & Giudici, P. (2024). How fair is machine learning in credit lending? *Quality and Reliability Engineering International*, 40(6), 3452–3464.
24. Chen, Y., Giudici, P., Liu, K., & Raffinetti, E. (2024). Measuring fairness in credit ratings. *Expert Systems with Applications*, 258, 125184.
25. Bahloul, R., Hewahi, N., & Elmedany, W. (2026). Performance, fairness, and explainability in AI-based credit scoring: A systematic literature review. *Journal of Risk and Financial Management*, 19(2), 104.
26. Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: An explainable novel hybrid machine learning approach. *Scientific Reports*, 14, 25174.
27. Tigges, M., Mestwerdt, S., Tschirner, S., & Mauer, R. (2024). Who gets the money? A qualitative analysis of fintech lending and credit scoring through the adoption of AI and alternative data. *Technological Forecasting and Social Change*, 205, 123491.
28. European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L 2024/1689.